UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

INTELLIGENT CYBERINFRASTRUCTURE FOR BIG DATA ENABLED HYDROLOGICAL MODELING, PREDICTION, AND EVALUATION

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

ZHANMING WAN
Norman, Oklahoma
2015

INTELLIGENT CYBERINFRASTRUCTURE FOR BIG DATA ENABLED HYDROLOGICAL MODELING, PREDICTION, AND EVALUATION

A DISSERTATION APPROVED FOR THE
GRADUATE COLLEGE

BY

_____
Dr. Yang Hong, Chair

_____
Dr. Aondover Tarhule

_____
Dr. Xiangming Xiao

_____
Dr. Valliappa Lakshmanan

_____
Dr. Sadiq Khan

_____
Dr. Naiyu Wang

This dissertation is dedicated to my wife Wei Sun, to my parents,

and to my parents-in-law, for their constant love, support, and patience.

Acknowledgements

I would like to express sincere appreciation to Dr. Yang Hong, my academic advisor, for his mentoring and support during the past three and half years of the doctoral degree program. I would also like to thank other members of my advisory committee, Dr. Aondover Tarhule, Dr. Xiangming Xiao, Dr. Valliappa Lakshmanan, Dr. Sadiq Khan, and Dr. Naiyu Wang, for all their help and advice through the development of my dissertation research. Special thanks to Dr. Jonathan J. Gourley, Dr. Ke Zhang and Dr. Xianwu Xue for their contribution during this work.

This work is mainly supported by Hydrometeorology and Remote Sensing (HyDROS) Lab and School of Civil Engineering and Environmental Science at the University of Oklahoma. Partial support was provided by National Oceanic and Atmospheric Administration (NOAA) and Advanced Radar Research Center (ARRC) at the University of Oklahoma.

Table of Contents

List of Tables

List of Figures

Abstract

Most hydrologic data are associated with spatiotemporal information, which is capable of presenting patterns and changes in both spatial and temporal aspects. The demands of retrieving, managing, analyzing, visualizing, and sharing these data have been continuously increasing. However, spatiotemporal hydrologic data are generally complex, which can be difficult to work with knowledge from hydrology alone. With the assistance of geographic information systems (GIS) and web-based technologies, a solution of establishing a cyberinfrastructure as the backbone to support such demands has emerged. This interdisciplinary dissertation described the advancement of traditional approaches for organizing and managing spatiotemporal hydrologic data, integrating and executing hydrologic models, analyzing and evaluating the results, and sharing the entire process.

A pilot study was conducted in Chapter 2, in which a globally shared flood cyberinfrastructure was created to collect, organize, and manage flood databases that visually provide useful information to authorities and the public in real-time. The cyberinfrastructure used public cloud services provided by Google Fusion Table and crowdsourcing data collection methods to provide location-based visualization as well as statistical analysis and graphing capabilities. This study intended to engage citizen-scientists and presented an opportunity to modernize the existing paradigm used to collect, manage, analyze, and visualize water-related disasters eventually.

An observationally based monthly evapotranspiration (ET) product was produced in Chapter 3, using the simple water balance equation across the conterminous United States (CONUS). The best quality ground- and satellite-based observations of the water

budget components, i.e., precipitation, runoff, and water storage change were adopted, while ET is computed as the residual. A land surface model-based downscaling approach to disaggregate the monthly GRACE equivalent water thickness (EWT) data to daily, 0.125º values was developed. The derived ET was evaluated against three sets of existing ET products and showed reliable results. The new ET product and the disaggregated GRACE data could be used as a benchmark dataset for researches in hydrological and climatological changes and terrestrial water and energy cycle dynamics over the CONUS.

The study in Chapter 4 developed an automated hydrological modeling framework for any non-hydrologists with internet access, who can organize hydrologic data, execute hydrologic models, and visualize results graphically and statistically for further analysis in real-time. By adopting Hadoop distributed file system (HDFS) and Apache Hive, the efficiency of data processing and query were significantly increased. Two lumped hydrologic models, lumped Coupled Routing and Excess STorage (CREST) model and HyMOD model, were integrated as a proof of concept in this web framework. Evaluation of selected basins over the CONUS were performed as a demonstration. Our vision is to simplify the processes of using hydrologic models for researchers and modelers, as well as to unlock the potential and educate the less experienced public on hydrologic models.

Chapter 1: Introduction

1.1 Statement of Problems

When conducting hydrologic research, various data may be acquired during the process. With the help of geographic information systems (GIS), the efforts of processing, analyzing, and visualizing the data is significantly diminished. However, several possible problems still exist. First, if the data is prepared by a third party, all we need to do after downloading the data is effectively organizing and using the data. Data is most commonly disseminated as data files. With file-based data structure, it is difficult to use and query the data, especially when the data volume is large. Second, if the required data is not provided by a centralized organization, it is difficult to collect these data by gathering limited labor to search and inquire. Third, data sharing and remote collaboration might be problematic with traditional stand-alone work environment.

Hydrological modeling is a widely adopted approach in hydrology research, which uses empirical, physical, or mathematical principles to conceptualize and simulate different phases of hydrologic cycle in order to assist scientific research in hydrologic events prediction, water resources management, and climate change assessment (McCuen 1973, Viessman and Lewis 2003, Xue et al. 2015). To execute a hydrologic model, one needs to install and configure the model, prepare all forcing data as model requires, and visualize and analyze the results. It is time-consuming to deal with these tedious steps. Moreover, it is difficult to share the whole process since it contains the model, the data, and the results.

Hydrologic models work as simulation of real hydrologic processes to predict hydrologic events. They can be effective if they provide accurate predictions. In fact, hydrologic models are not always accurate; even if they perform well for one basin, there is no guarantee they can perform well for another basin. Calibration of hydrologic models is one method to improve the performance of the models by using optimization algorithms to adjust model parameters in given ranges to amend the results being close to the observations. However, most of the times model results still differ from the observations due to a combination of reasons, including spatial variabilities and uncertainty of basins, data, and models.

Therefore, it is challenging but timely to establish a cyberinfrastructure for processing and organizing hydrologic data, setting up and configuring hydrologic models, analyzing and visualizing both the data and the results, and sharing the entire process and collaborating with people, including non-hydrologist as well as researchers around the world.

1.2 Related Background

1.2.1 Hydrologic Data

Most hydrologic data have spatial and temporal attributes, which can be illustrated in different forms. Vector shapes, including points, polylines, and polygons, are commonly used as the media to demonstrate spatial attribute of such data. For instance, points can be used to indicate locations of events of natural hazards, such as floods (Adhikari et al. 2010, Wan et al. 2014) as used in Chapter 2 and landslides (Li, Liu, et al. 2013), and locations of ground gauge stations (Gourley et al. 2013, Wan et al. 2015). Polylines can be used to show locations and shapes of rivers and contour lines (Viessman and Lewis 2003). Polygons are widely used in inundation mapping (Fluet-Chouinard et al. 2015) and watersheds delineation (Martínez-López et al. 2014). Besides vector data, raster data serve as an alternative to represent data with spatial attribute. The raster data structure, typically referred to as grid or matrix, comprises an array of pixels (i.e. grid cells) in predefined orders with coordinate information for each pixel (Mattikalli 1995). Many satellite observations used in hydrology research are conveyed as raster data, such as the Gravity Recovery and Climate Experiment (GRACE) data used in Chapter 3. Likewise, algorithm generated hydrologic data have a similar format, such as observation interpolation based Parameter-elevation Regressions on Independent Slopes Model (PRISM) data used in Chapters 3 and 4, and calculated ET products used and reconstructed in Chapter 3.

1.2.2 Hydrologic Models

Hydrologic models are software packages used to simulate single or multiple surface and underground hydrologic processes in hydrologic cycle to understand such processes by recreating historical events and predicting potential future occurrences (Viessman and Lewis 2003). By determining if the domain can be subdivided, hydrologic models are categorized into lumped, semi-distributed or distributed models, while in semi-distributed and distributed models the target basin can be subdivided into sublevel computational units using mathematical or physical representations (Vieux 2001). Semi-distributed and distributed hydrologic models are capable of maintaining spatial variabilities of basins. Thus, model complexity is increased, which influences model efficiency when applying forcing data with relatively higher resolutions (Carpenter and Georgakakos 2006). In addition, model calibration is utilized to optimize model parameters and obtain satisfactory results from hydrologic models. In Chapter 3, soil moisture content data are simulated by four land surface / hydrologic models. The research in Chapter 4 integrates two lumped hydrologic models for rainfall-runoff simulations.

1.2.3 GIS and Web Integrations

As aforementioned, hydrologic data are mostly spatiotemporal data in both vector and raster formats. Therefore, desktop and web-based GIS are being heavily used for data processing, analysis, visualization, and sharing (Bhatt, Kumar, and Duffy 2014, DeVantier and Feldman 1993). Some pilot studies have been conducted in hydrology using desktop GIS (e.g. ArcGIS and Quantum GIS (QGIS)). Zhan and Huang (2004)

developed an extension of ArcGIS to determine curve numbers (CN) from soil and land use information and calculate runoff using Soil Conservation Service (SCS) CN procedure. Olivera et al. (2006) presented a combination of ArcGIS with the Soil and Water Assessment Tool (SWAT) for data analysis and visualization. Bhatt, Kumar, and Duffy (2014) introduced a modeling framework coupled Quantum GIS (QGIS) with Penn State Integrated Hydrologic Model (PSIHM) for watershed delineation, simulation, and visualization.

With fast Internet access and low latency, traditional read-only web environment has been transformed into the "rich application" era of Web 2.0. Plenty of data transfer protocols, file formats, open source software, and online services are developed to utilize the bandwidth of the Internet and provide novel user experience. HyperText Markup Language (HTML) is the basic language to create web pages. Just like HTML, Extensible Markup Language (XML), JavaScript Object Notation (JSON) and GeoJSON are developed to encoding regular data and spatial data in formats that can be read by both human and machine efficiently. With asynchronous methods (e.g. Asynchronous JavaScript and XML (AJAX)) to get and post data in standard formats over the internet, data becomes sharable and concurrency of access from multiple users becomes feasible. The development of web environment and web technology enable the capability to share hydrological modeling framework and collaborate with people distantly, Huang (2003) integrated TOPography based hydrological MODEL (TOPMODEL) into a web modeling and visualization framework to allow user interaction with environmental applications. The famous Web-based Hydrograph Analysis Tool (WHAT) is an example using GIS to access real-time U.S. Geological Survey (USGS) discharge data directly from web

servers for visualization and validation (Lim et al. 2005). Comair et al. (2014) presented a GIS based hydrologic information system to store and share hydrologic data, execute hydrologic model, and deliver simulation results for better water management and decision-making.

1.2.4 Cloud Computing and Big Data Era

Beyond traditional web technology, the use of cloud computing and big data has surged in recent years to facilitate research in geo-computation and hydrology. Cloud computing is a framework for supporting elastic network access to elastic computing resources (Mell and Grance 2011) so that it can maximize the efficiency and data safety during collaboration, while minimizing time and expense spent on the system. Several studies have confirmed the advantages of using cloud computing. Gong, Yue, and Zhou (2010) adopted cloud computing to provide elastic geoprocessing capabilities and data services in a distributed environment. Behzad et al. (2011) used cloud computing with cyber-infrastructure-based GIS to implement a large number of concurrent groundwater ensemble runs with improvement in computational efficiency. The study of Sun (2013) presented a collaborative decision-making water management system using a cloud-based service, revealing the potential to fundamentally change a water management system from its design to the operation. Namibia flood SensorWeb infrastructure is another example which was created for rapid acquisition and distribution of data products for decision-support systems to monitor floods, utilizing the Matsu Cloud to store and pre-process data through hydrological models and eliminating the latency to users' real-time online requests (Kussul et al. 2012).

Big data refers to any collection of data sets with large volume, high complexity, and fast update rate, which are usually difficult to be managed and processed by traditional databases or file systems (Kitchin 2013, Bhosale and Gadekar 2014, Fernández et al. 2014, Vitolo et al. 2015). The advancement of satellite and ground observation technologies improve both spatial and temporal resolutions recent years, which in return transforms spatiotemporal data in earth science studies into big data scope. With big data being involved in hydrological modeling, solutions such as distributed file systems and high-performance computation are incorporated into such framework. Li, Yang, et al. (2013) developed a web-based application for high-performance analysis and visualization of big spatiotemporal data and climate model simulations. Hu, Cai, and DuPont (2015) coupled an agent-based system with an environmental model and integrated it into a web application enabled by Hadoop-based high-performance computation for watershed management.

1.2.5 Data Visualization

The increase in computation power is always associated with the development of web-based visualization. Improved graphic processing abilities offer two dimensional (2D), three dimensional (3D), and even multi-dimensional options for complex and high-resolution geospatial data visualization for web environment. Several well-known examples include ArcGIS Online, a GIS solution for web-based analysis, visualization and collaboration (ESRI 2015); Earth, an online 3D animated visualization of global weather conditions (Earth 2015); and Google Earth, a 3D geographic information program for both desktop and online uses (Google 2015a). When Al Gore mentioned

computational technology and digital earth concept in his speech in 1998 (Gore 1998), one could hardly imagine what has been achieved today in this field.

From programming perspective, Application Programming Interfaces (API) are provided for both client side and server side. By using libraries like Google Maps API (Google 2015c) and Leaflet (Leaflet 2015), it is straightforward to create simple customized applications with map visualization. Open data kit (ODK) (ODK 2015) is another example of open source software to help build data collection survey, collect data from users, and aggregate data into useful formats.

1.3 Research Objectives

The overarching objective of this dissertation is to establish an intelligent cyberinfrastructure for organization and management of hydrologic data, integration and execution of hydrologic models, analysis and evaluation of the results, and sharing the entire process. As an interdisciplinary research, the vision is to facilitate the existing paradigm of hydrologic research and expose potential of the water community by engaging the public (including non-hydrologist) with state-of-the-art web technologies and the assistance of GIS to help with hydrologic data collection, use hydrological modeling, and conduct hydrologic analytics. Specific objectives in this dissertation are listed as follows:

- To organize natural hazards inventory by establishing a cloud-based cyberinfrastructure with capabilities of on demand data management and sharing, instantaneous map and statistic visualizations, and real-time updating through crowdsourcing.
- To develop land surface models-based algorithm to disaggregate GRACE data to improve and produce an important model output (ET) using water balance-based approach.
- To establish a web GIS-based hydrological modeling framework with big data support and scalable structure to facilitate multi-source data processing, evaluation, and visualization.

1.4 Organization of the Dissertation

This dissertation is organized into an introductory chapter, three main chapters with three related researches, and a summary chapter. Chapters 2 and 3 were published as two stand-alone peer-reviewed journal articles, and Chapter 4 is ready to be submitted to a journal. List of publications from this dissertation is in Section 1.5.

Chapter 2 presents a study entitled “A cloud-based global flood disaster community cyber-infrastructure: Development and demonstration”. A globally shared flood cyber-infrastructure to collect, organize, and manage flood databases that visually provide useful information in real-time, using cloud computing services and crowdsourcing data collection methods to provide on-demand, location-based visualization and statistical analysis. It involves public participation to submit their entries of flood events for archiving comprehensive information of flood events, past and present. As a cloud-based cyberinfrastructure with crowdsourcing capabilities, an opportunity is presented to modernize the way of collecting and sharing information for water related disasters.

Chapter 3 is entitled “Water balance-based actual evapotranspiration reconstruction from ground and satellite observations over the conterminous United States”. This chapter describes an approach to produce an observationally based monthly ET product using the water balance equation across the CONUS. The best quality ground- and satellite-based observations of the water budget components are adopted, while ET is computed as the residual. A land surface models-based downscaling approach to disaggregate the GRACE EWT data is developed. The reconstructed ET is evaluated against three sets of existing ET products, showing similar spatial patterns and small

differences that ensure the reliability of this approach. The new ET product can be used as a benchmark dataset in evaluation for other hydro-climatological research over the CONUS.

Chapter 4 is entitled “Big data solutions enabled web GIS-based hydrological modeling framework for the conterminous United States”. It elaborates on an automated general-purpose hydrological modeling framework with web accessibility for non-hydrologists to organize hydrologic data, execute hydrologic models, and visualize results graphically and statistically for further analysis in real-time. Aided by HDFS and Apache Hive, the framework presents an efficient and effective way for data processing and query. Two lumped hydrologic models, lumped CREST and HyMOD, were integrated as a proof of concept in this web GIS framework and it is evaluated against selected basins over the CONUS. The goal is to simplify the processes of using hydrologic models for researchers and modelers, and educate the less experienced non-hydrologist on hydrologic models. More importantly, this shared framework is designed with elasticity of being expanded to accommodate various data and different models.

1.5 List of Publications from the Dissertation

Chapter 2

Wan, Zhanming, Yang Hong, Sadiq Khan, Jonathan Gourley, Zachary Flamig, Dalia Kirschbaum, and Guoqiang Tang. 2014. "A cloud-based global flood disaster community cyber-infrastructure: Development and demonstration." Environmental Modelling & Software 58:86-94.

Chapter 3

Wan, Zhanming, Ke Zhang, Xianwu Xue, Zhen Hong, Yang Hong, and Jonathan J Gourley. 2015. "Water balance‐based actual evapotranspiration reconstruction from ground and satellite observations over the conterminous United States." Water Resources Research 51 (8):6485-6499.

Chapter 4

Wan, Zhanming, Xianwu Xue, Ke Zhang, Yang Hong, Jonathan Gourley, and Humberto Vergara. 2016. "Big data solutions enabled web GIS-based hydrological modeling framework for the conterminous United States." (To be submitted to Environmental Modelling & Software).

Chapter 2: A cloud-based global flood disaster community cyber-infrastructure: Development and demonstration

Flood disasters have significant impacts on the development of communities globally, often causing loss of life and property. It is increasingly important to create a globally shared flood cyber-infrastructure (CyberFlood) to collect, organize, and manage flood databases that visually provide useful information back to both authorities and the public in real-time. The community shared CyberFlood infrastructure described in this study uses cloud computing services and crowdsourcing data collection methods to provide on-demand, location-based visualization as well as statistical analysis and graphing capabilities. It also involves public participation, allowing the public to submit their entries of flood events to help the community to archive comprehensive information of flood events, past and present. The Global Flood Inventory (GFI) is used as a primary database to develop this cyber-infrastructure. The GFI, which contains detailed information of global flood events from 1998 to 2008, was developed and made available for community use. In order to expand and update the existing inventory, a crowdsourcing methodology is employed which enables web-based data entry for the public to report or record their personal accounts of local flood events. This step is also intended to engage citizen-scientists so that they may become motivated and educated about the latest developments in satellite remote sensing and hydrological modeling technologies. Cloud computing is further integrated into this cyber-infrastructure by utilizing public cloud services provided by Google, which effectively accelerates the speed during data processing and visualization over the Internet. As a cloud-based cyber-infrastructure,

people can access this infrastructure from all over the world through the Internet or mobile phones. The shared vision is to better serve the global water community by providing essential flood information, aided by the state-of-the-art cloud computing and crowd-sourcing technology. This CyberFlood presents an opportunity to eventually modernize the existing paradigm used to collect, manage, analyze, and visualize water-related disasters (e.g. floods, landslide, and droughts).

2.1 Introduction

Flooding is one of the most dangerous natural disasters globally, frequently causing tremendous loss of life and economic damages. According to the International Federation of Red Cross (IFRC) and Red Crescent Societies (RCS), almost half of the natural disasters that happened between 2002 and 2011 were floods. During this period, natural disasters caused approximately 1.1 million fatalities worldwide, affected approximately 2.7 billion people, and led to economic losses totaling approximately $1.4 trillion USD. Of these damages, approximately 57,000 (5%) of the fatalities, 1.2 billion (44%) of the affected, and $278 billion USD (20%) of the economic damages were attributed to floods alone (Zetter 2012).

The significant global impact of recurring flooding events leads to an increased demand to have comprehensive flood databases for flood hazard studies. There are several existing flood databases, such as the International Disaster Database (EM-DAT), ReliefWeb (launched by the United Nations Office for the Coordination of Humanitarian Affairs (OCHA)), the International Flood Network (IFNET) and the Global Active Archive of Large Flood Events (created by the Dartmouth Flood Observatory (DFO)). However, there is often a lack of specific geospatial characteristics of the flooding impacts or a failure to enlist all flood events due to variable entry criteria. Moreover, these data warehouses lack interactive information sharing with the communities affected by the flood events. Therefore, a methodology developed by Adhikari et al. (2010) utilized valuable flood event information from the aforementioned sources, specifically the DFO, and synthesized these data with media reports and remote sensing imagery in order to provide a record of flooding events from 1998 to 2008. The digitized Global Flood

Inventory (GFI) gathers and organizes detailed information of flood events from reliable data sources, defines and standardizes categorical terms as entry criteria for flood events (e.g. severity and cause), and cross-checks and quality controls flood event information (e.g. location) to eliminate redundant listings. These characteristics make GFI an appropriate starting point to develop a unified, global flood cyber-infrastructure. However, one limitation of this database is that GFI only contains flood events through 2008. Although it is possible that flood events after 2008 can be collected manually, as was done in Adhikari et al. (2010), it can be incomplete and inefficient since this process only involves a limited number of resources and people. Recently, technological advances in social media have tremendously improved data gathering and dissemination, especially with the development of World Wide Web technologies. Built on the platform of social media, crowdsourcing has become a versatile act of collecting information from the public.

Crowdsourcing is a term that generally refers to methods of data creation, where large groups of potential individuals generate content as a solution of a certain problem for the crowdsourcing initiator(Hudson-Smith et al. 2009, Estellés-Arolas and González-Ladrón-de-Guevara 2012). In theory, crowdsourcing is based on two assumptions described by Goodchild and Glennon (2010). First, “a group can solve a problem more effectively than an expert, despite the group’s lack of relevant expertise”, and second, “information obtained from a crowd of many observers is likely to be closer to the truth than information obtained from one observer.” Based on the definition and assumption of crowdsourcing, it has the ability to collect a considerable amount of information from its randomly distributed participants. The nature of crowdsourcing accommodates data

collection in numerous forms, including questionnaires, phone calls, text messages, emails, web surveys and other paper-based, mobile phone-based, and web-based methods. Moreover, crowdsourcing can be embedded with location-based information by using GPS-enabled devices, IP (internet protocol) addresses, or participants' awareness of their current locations. Crowdsourcing offers new opportunities to expand the information available to impacted communities and provide a "two-way" street for the same affected populations to communicate with the global community.

The data collected from crowdsourcing will be used in a cloud computing framework for information sharing that includes data processing and visualization. Gong, Yue, and Zhou (2010) adopted cloud computing technology in geoprocessing functions to provide elastic geoprocessing capabilities and data services in a distributed environment. Behzad et al. (2011) used cloud computing in addition to a cyber-infrastructure-based geographic information system to facilitate a large number of concurrent groundwater ensemble runs by improving computational efficiency. Huang et al. (2013) integrated cloud computing in dust storm forecasting to support scalable computing resources management, high resolution forecasting, and massive concurrent computing. As defined by the National Institute of Standards and Technology (NIST), cloud computing is a model for supporting elastic network access to a shared pool of configurable computing resources (Mell and Grance 2011). The nature of cloud computing assures that it can (a) reduce the time and cost during implementation, operation, and maintenance of the global flood cyber-infrastructure, (b) provide an interface for collaboration at both global and local scales, and (c) conveniently share data in a secure environment. These advantages make cloud computing an attractive technique

in the global flood cyber-infrastructure that can maximize the efficiency and data safety during collaboration, while minimizing time and expense spent on the system.

Several studies have already used cloud-based services with water-related management and monitoring. The study of Sun (2013) presented a collaborative decision-making water management system using a cloud service provided by Google Fusion Table. The author describes the migration of the management system from a traditional client-server-based architecture to a cloud-based web system, revealing the potential to fundamentally change a water management system from its design to the operation. Another example is the Namibia flood SensorWeb infrastructure, which was created for rapid acquisition and distribution of data products for decision-support systems to monitor floods (Kussul et al. 2012). The decision-support system utilizes the Matsu Cloud to store and pre-process data through hydrological models, eliminating the latency when clients select specific data.

This study proposes a cloud-computing service provided by Google to establish the global flood cyber-infrastructure, to share the GFI, to provide statistical and graphical visualizations of the data, and to expand the breadth and content of the GFI by collecting new flood data using crowdsourcing technology (i.e. CyberFlood). The next section focuses on the architecture of the cloud computing system designed for global flood monitoring, analysis, and reporting. Section 2.3 demonstrates the system's functionality, and a summary is provided in Section 2.4.

2.2 Cyber-infrastructure Design for Flood Monitoring

The global flood cyber-infrastructure consists of four components: the GFI data source, cloud service, web server, and client interface (Figure 2.1). The GFI is pre-processed before being imported into the cyber-infrastructure, as explained later in this section. The cloud service, which significantly improves the performance and decreases the burden on the web server, handles all data queries, data visualization, and data analysis. The web server simply deals with sending requests and responses between clients and the cloud. The client interface is mainly built with hypertext markup language (HTML) and JavaScript. Since all the data are processed before being imported into this cyber-infrastructure, the client side only sends operational requests from users and renders responses from the cloud service.

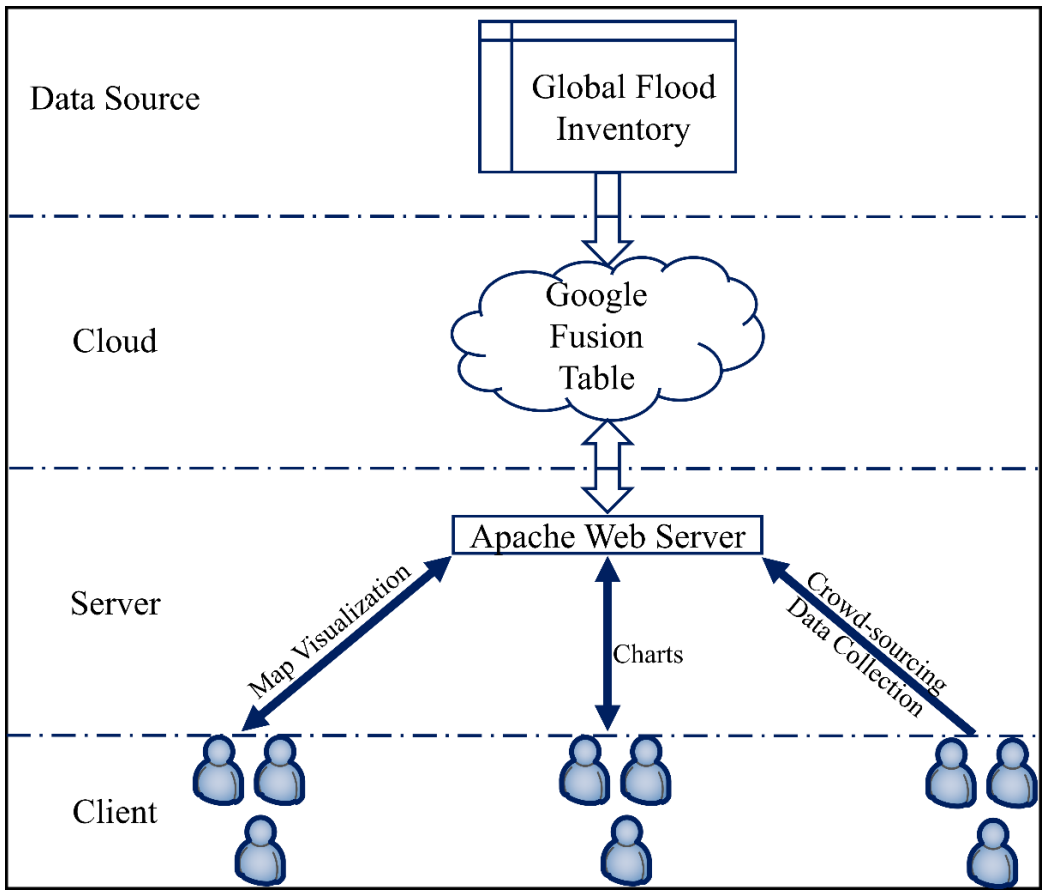


Figure 2.1 The global flood community cyber-infrastructure framework.

As previously mentioned, GFI standardizes categorical terms as entry criteria for flood events. In other words, every data column in GFI is carefully designed so that each entry strictly follows the criteria of the corresponding data column (Figure 2.2a). GFI was pre-processed before being successfully imported into a Google Fusion Table. Python code, which is a cross-platform, extensible, and scalable programming language (Sanner 1999), was written to do the data conversion. The purpose is to maintain data consistency, making the converted data readily readable and reducing the data conversion load on the client side. In this process, cells containing -9999, which represent no value in GFI, are removed because they are not consistent with empty cells that also represent no value. Data columns of flood severity, cause, country, and continent are filled with numbers to

indicate certain meanings in GFI. A look-up table was used to convert the numerical codes into text. For example, “1” means “heavy rain” in the column pertaining to flood causes, whereas it means “Africa” in the column pertaining to continents (Figure 2.2a and Figure 2.2b). In other words, if the GFI with numbers are imported into the Fusion Table and used directly by the cyber-infrastructure, the numbers have to be converted to the corresponding texts each time during the refresh on the client side. As a result, text is assigned to severity, cause, country, and continent during this process. Location, the most important information for map visualization in this cyber-infrastructure, is described in two columns representing latitude and longitude in GFI. However, if one flood event involves more than one location, then there will be multiple data records, and only the first data record has shared information such as event ID and date. To improve this data structure and for better visualization, multiple data records representing the same flood event are combined into a single record, while location is presented as MultiGeometry using Keyhole Markup Language (KML) (Wilson 2008).

ID	Year	Month	Day	Duration	fatality	Severity	Cause	Lat	Long	Country code	Continent Code
2707	2008	12	28	23	25	1	2, 1	-22.92	34.03	140	1
2706	2008	12	26	18	24	1	1	-3.33	103.14	93	3
2705	2008	12	26	3	-9999	1	1	44.66	-123.53	213	6
2704	2008	12	26	3	-9999	1	1	41.04	-89.46	213	6
2703	2008	12	25	12	9	1	1	16.89	107.06	219	3
2702	2008	12	13	31	76	1.5	1	9	-74.23	42	8
2701	2008	12	13	2	2	1	1	51.49	-1.73	212	5

a. Global Flood Inventory Data Table

ID	Year	Month	Date	Duration	Fatality	Severity	Cause	Geometry	CountryCode	ContinentCode
2707	2008	12	12/28/2008	23	25	Class 1	Tropical cyclone, Heavy rain	-22.92,34.03	Mozambique	Africa
2706	2008	12	12/26/2008	18	24	Class 1	Heavy rain	-3.33,103.14	Indonesia	South East Asia
2705	2008	12	12/26/2008	3		Class 1	Heavy rain	44.66,-123.53	United States	North America
2704	2008	12	12/26/2008	3		Class 1	Heavy rain	41.04,-89.46	United States	North America
2703	2008	12	12/25/2008	12	9	Class 1	Heavy rain	16.89,107.06	Vietnam	South East Asia
2702	2008	12	12/13/2008	31	76	Class 2	Heavy rain	9,-74.23	Colombia	South America
2701	2008	12	12/13/2008	2	2	Class 1	Heavy rain	51.49,-1.73	United Kingdom	Europe

b. Google Fusion Table

Figure 2.2 Comparison of data tables a) global flood inventory and b) Google fusion table.

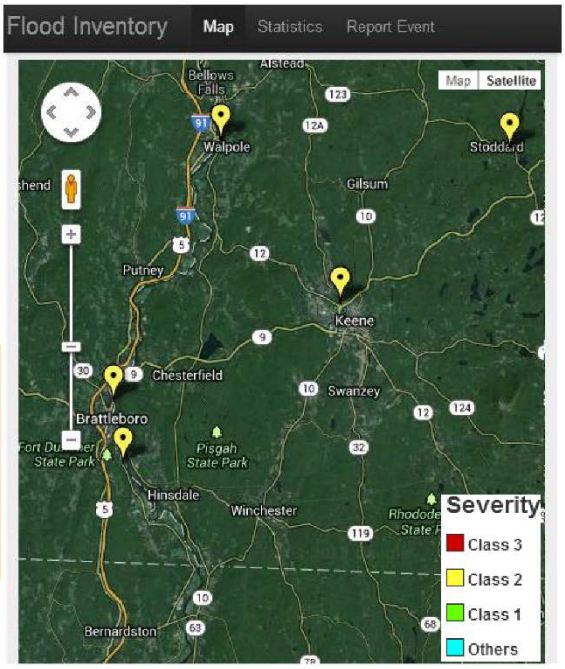
An example of a flooding event in New Hampshire in October 2005 is illustrated in Figure 2.3. Figure 2.3a shows the event as stored in the original GFI covering events from 1998-2008. Five locations were associated with this event. Cells are left blank if they share the same record as in the first row. Figure 2.3b shows the same flooding event as in Figure 2.3a, but converted into a Google Fusion Table. This table also includes all five locations that are now represented in the geometry column with KML. Figure 2.3c illustrates the visualization of this event, showing the severity as well as the specific locations impacted. Additional layers such as rivers, roads, and topography can also be included during this step to ascertain the spatial extent of inundation.

ID	Year	Month	Day	Duration	fatality	Severity	Cause	Lat	Long	Country code	Continent Code
1859	2005	10	8	10	11	1.5	1	42.9475	-72.2944	213	6
								43.07667	-72.0989		
								43.08389	-72.4317		
								42.86528	-72.555		
								42.8125	-72.5444		

a. Global Flood Inventory

ID	1859
Year	2005
Month	10
Date	10/08/2005
Duration	10
Fatality	11
Severity	Class 2
Cause	Heavy rain
Geometry	<Point><coordinates>-72.29444444,42.9475</coordinates></Point><Point><coordinates>-72.09888889,43.07666667</coordinates></Point><Point><coordinates>-72.43166667,43.08388889</coordinates></Point><Point><coordinates>-72.555,42.86527778</coordinates></Point><Point><coordinates>-72.54444444,42.8125</coordinates></Point>
CountryCode	United States
ContinentCode	North America

b. Google Fusion Table



c. Google Map View

Figure 2.3 Flood event over Northeast U.S. in New Hampshire of October 2005 a) global flood inventory, b) Google fusion table attributes, and c) Google map view.

The processed GFI, now converted to a Google Fusion Table (Figure 2.2b) belongs to a "Software as a Service" (Yang et al. 2011) type of cloud-based service for data management and integration (Gonzalez et al. 2010). Google Fusion Table was created to manage and collaborate with tabular datasets in which geospatial fields can be included to provide location information. These geospatial fields can be in the form of latitude and longitude in two separate columns, latitude and longitude pairs in one column, or KML strings in one column. Fusion Table accepts many different tabular formats of files as its data source. Any text-delimited files such as comma-separated values (CSV) files, KML files, and spreadsheets can be imported directly into a Fusion Table. Since

Google Fusion Table is a part of Google Drive, users can simply select an existing spreadsheet from their Google Drive and import it into a Fusion Table. Cloud computing is embedded to provide rapid responses to requests from users for data querying, summary, and visualization. Moreover, data security and sharing is already implemented in Google Fusion Table.

The steps required to import data into a Google Fusion Table are straightforward. First, the data must be in one of the supported formats (tabular or text-delimited data such as CSV files, excel spreadsheets, and other similar types.). A wizard then provides easy-to-follow instructions describing how to upload the data. Fusion Table looks like a common table in a spreadsheet, whereas it supports structured query language (SQL) to operate the table. Keywords, such as “SELECT”, “INSERT”, “DELETE” and “UPDATE”, can be used to manipulate Fusion Table, which is similar to how a table is handled in a database. Fusion Table provides application programming interface (API) to programmatically perform SQL-based, table-related tasks through using hypertext transfer protocol (HTTP) requests (Google 2015b). By combining with other Google-provided APIs, the capability of Fusion Table can be extended to not only manipulate the data in the table, but also visualize the data through thematic mapping and analytic charts.

Fusion Table, which plays an important role in this global flood cyber-infrastructure, provides data storage, data sharing, and fast data access. However, since the infrastructure is functioning from the backend, users cannot benefit from this service unless a traditional server and client components are included for interaction. Since all the computing loads are on the cloud, the web server only serves as a “middleware” dealing with requests and responses between the cloud and clients. The web server also

protects the Fusion Table on the cloud from being accidentally modified by clients. Google provides two kinds of API keys for programmers to develop applications. One of the keys is a string, which grants permission to applications to select items from the Fusion Table. The other key is a special file that should be stored securely with the application on the web server. This type of key grants permission to the application from the specific web server to insert, update, or delete items from the Fusion Table. The client side is programmed with HTML and JavaScript, along with several APIs from Google, to send requests through the server to the cloud, receiving responses for location-based and analytic visualization.

2.3 Demonstration

The global flood cyber-infrastructure is currently running at http://eos.ou.edu/flood/ (Figure 2.4). An Apache web server is deployed to host the frontend web interface. Google Map has been integrated to map the locations of flood events after querying the Fusion Table using the Google Map API. All the points representing locations of flood events are color coded by severity or fatalities associated to the flood event. Severity is classified into classes 1, 2, and 3, with “Class 1” being least severe and “Class 3” the most severe. Fatalities are categorized into four classes based on the value. Users are allowed to select a range of years and causes of flood events from the provided controls. Each selection will lead to a new query from the Fusion Table, which means that the desired data will be plotted on the map with event details that have just been uploaded in real-time. In addition to visualization of the data using information stored in the Fusion Table, a Google Chart API is utilized to create analytic charts for statistical analysis of the flood events (Figure 2.5). Variables such as the year, month, severity, cause, continent, and country, can be analyzed in a chart and a table. Variables can be summarized by the count of the variables, sum of fatalities, or average of fatalities. For instance, Figure 2.5 demonstrates the summary of flood events by year and severity. Flood events with Class 1 severity are in a blue color on the chart, with about 270 of the flood events in 2003 occurring with such a severity class.

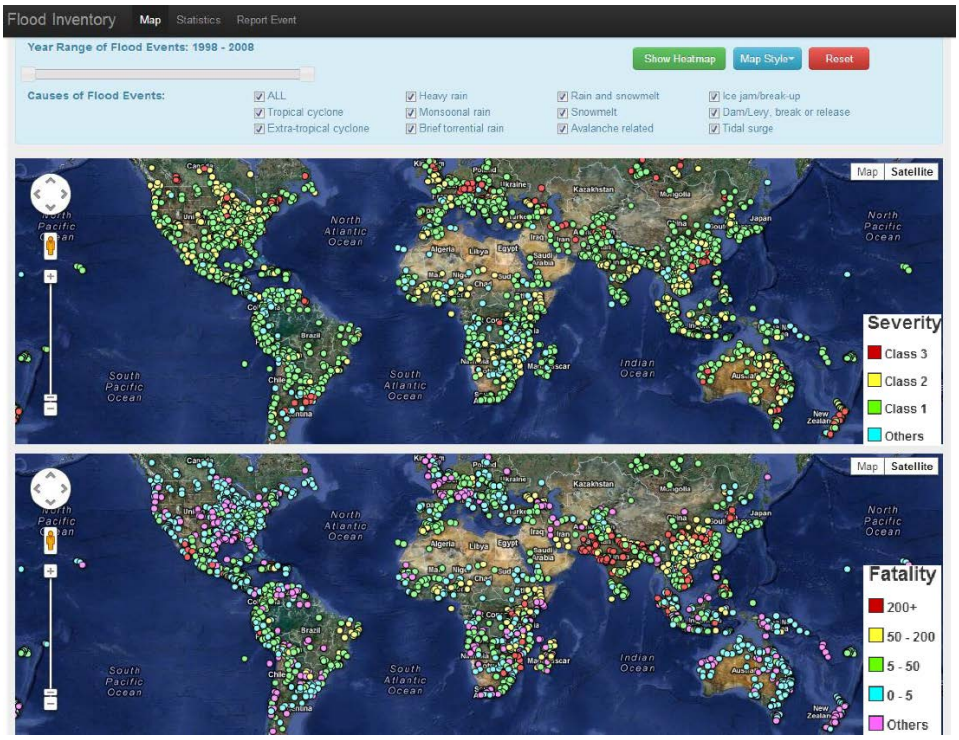


Figure 2.4 The map visualization of global flood cyber-infrastructure. The top and bottom maps are color coded by severity and fatalities respectively.

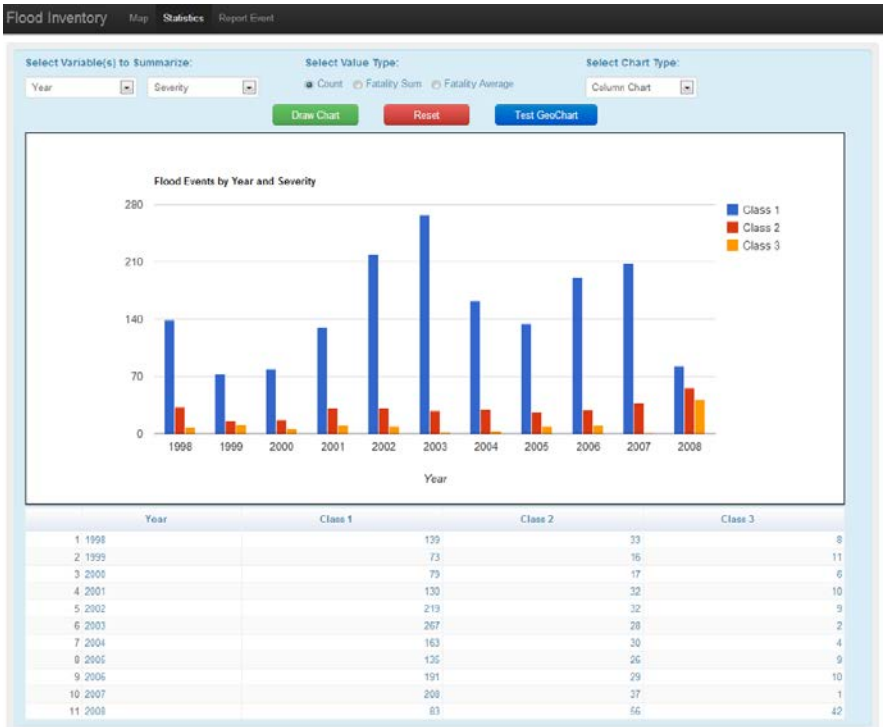


Figure 2.5 The statistical chart and table of global flood cyber-infrastructure.

In order to expand and update the existing GFI, now stored as a Fusion Table, crowdsourcing from public entries is implemented in this cyber-infrastructure by providing a flood events observation report form (Figure 2.6). Most of the fields are the same as the existing GFI. However, photo URL and source URL fields are appended to the Fusion Table to store additional details about the submitted flood event. This means that users are able to upload one photo per submission and provide a URL of the web source as a proof or supplemental information of that flood event. The current date will be retrieved from the users' operating system by default to submit present flood events. Users can also select any date between 1998 to present if past events are reported. Since reported events will be displayed on the map in real-time immediately following submission, location is a required field in the report form. Location will be automatically retrieved if a location service is allowed by the client's browser or the uploaded photo is geo-tagged. This report form is submitted directly into the Fusion Table through the server, and this process is protected by Google Account Authentication and Authorization Mechanism to secure data on the Fusion Table. A two-way quality control approach of data from crowdsourcing is implemented. First, when a user submits a report of flood events, the system will automatically check if each field is correctly formatted. For example, fields of latitude and longitude can only be numeric values. Fields of day, month, and year are restricted to certain numbers that can only be selected by users. Instructions have also been created for first-time users and they can learn what each field means and how to retrieve current location to help them submit correct information. Secondly, after submission, the data will be manually checked with different sources, including news reports, flood reports from other major disaster data sources, and satellite imagery.

Checking data sequentially is not an efficient way of quality control. However, it is effective in this case since the number of data received so far is limited. Newly submitted events following post-processing will be assigned IDs according to the number of milliseconds from 1970/01/01 to the time of the submission. For example, a flood event reported at 12/18/2013 23:35:15.199 will be assigned an ID of 1387431315199. Sequential IDs will be assigned to newly submitted data after quality control is complete. If crowdsourced data submissions increase in frequency in the future, automated data quality control procedures will be developed to check the spatial and temporal consistency with other flood reports. Other automated procedures can crosscheck the reports with global flood forecasts available from http://eos.ou.edu/Global_Flood.html. A crowdsourcing way to control the quality of crowdsourced flood events reports are under consideration. A mechanism could be established to grant permission to qualified users and students who have expertise in flood monitoring and validation to check the data quality in the Fusion Table.

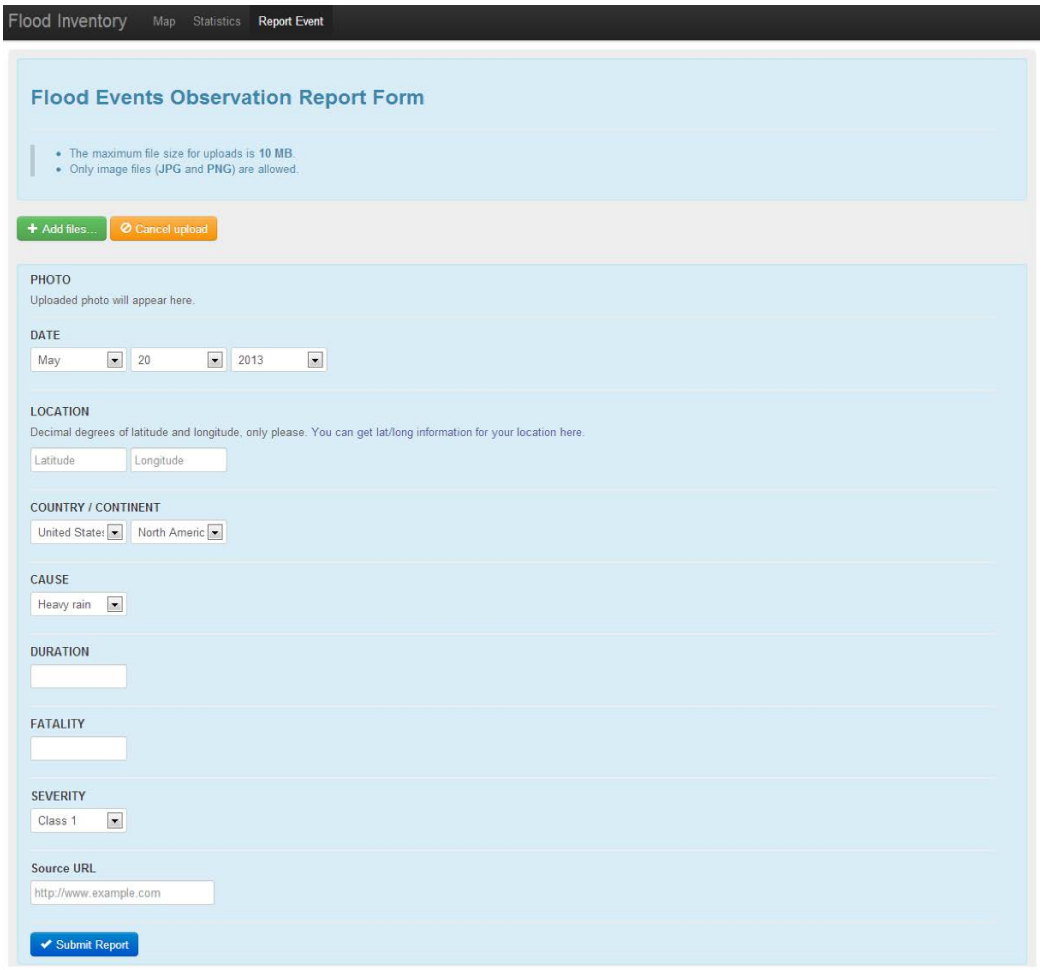


Figure 2.6 The flood events observation report form.

2.4 Discussion

2.4.1 Advantage

Although CyberFlood does not directly solve flooding problems, this work is expected to be able to help advance flood-related research areas such as hydrologic model evaluation, flood risk management, and flood awareness. Both the public and research community can use the resources provided by this cyber-infrastructure to analyze retrospective flood events and submit their witness accounts of previously unreported flood events. Therefore, this approach is useful for flood monitoring and validation research. The long-term database could also help generate flood climatology of occurrences and damage and therefore could potentially lead to better flood risk management for zoning and other flood-related decision-making purposes. Public engagement using crowdsourcing and cloud-based techniques could potentially raise flood awareness around the globe and provoke citizen-scientists to consider careers in the natural sciences, engineering, and mathematics.

CyberFlood has been created to be used by anyone with internet access. In order to access the flood resources, a web-based interface is provided and is becoming accessible through iOS apps for mobile users. As CyberFlood becomes more accessible through these apps, more people will use it to view retrospective flood events, monitor current flood events, and contribute to the flood community by submitting their reports of flood events. CyberFlood has been created to adapt the idea of Volunteered Geographic Information (VGI), which is described as tools to create, assemble, and disseminate geographic information provided voluntarily by individuals (Goodchild, Yuan, and Cova

2007), for compiling flood events by involving map-based visualization and utilizing human sensors to collect useful data globally.

Compared with the traditional server-client structure, the cloud computing service provided by Google Fusion Table enhances the performance of the global flood cyber-infrastructure in terms of the speed during data query and data visualization. By providing a Fusion Table API, the complexity of the global flood cyber-infrastructure is significantly reduced. This benefits both programmers and clients since they are able to focus more on the actual functions they need to implement and use, not on the logistics with the cloud itself. Rather than using the traditional server-client based structure, this simplified cloud-based framework makes it easier to develop scalable applications. Furthermore, taking into consideration of data sharing and collaboration, Fusion Table provides a comprehensive solution to keep data secure while making seamless communications between collaborators and Google servers for data updates, queries, and visualization.

2.4.2 Performance Experiment

An experiment was developed to compare the speed of reading data and geographically displaying data using Google Maps API with a Google Fusion Table and a MySQL database respectively, both of which contain the same dataset. Google Maps API provides two ways to display markers on Google Map. The traditional way is by using google.maps.Marker class. The more efficient way is to utilize google.maps.FusionTablesLayer class that can only be employed by data from the Google Fusion Table. As a result, the data in the Google Fusion Table is visualized by

google.maps.FusionTablesLayer class while the data in the MySQL database is visualized by google.maps.Marker class in this experiment. The query speed of both Google Fusion Table and MySQL database are rapid, taking a few milliseconds. However, the speed advantage becomes predominant when using data from the Google Fusion Table with google.maps.FusionTablesLayer class. Table 2.1 demonstrates the results of this performance experiment. The first 1000, 5000, 10000, 50000, and 100000 records are retrieved from the dataset. The average time of reading and displaying different size of data is calculated from five consecutive measurements. When data records increase from 1000 to 100000, the average elapsed time for using the Google Fusion Table with google.maps.FusionTablesLayer class is always low (less than 10 ms) while the average elapsed time for using the MySQL database with google.maps.Marker class is much higher (more than 1000 ms) and increases significantly to more than 3000 ms when displaying 100000 records.

Table 2.1 Performance comparison results.

	Google Fusion Table (google.maps.FusionTablesLayer)					
Test Order	**1**	**2**	**3**	**4**	**5**	**Average**
1,000 Records	17	8	8	7	8	9.6
5,000 Records	9	7	6	7	9	7.6
10,000 Records	12	7	8	6	7	8.0
50,000 Records	8	9	8	7	7	7.8
100,000 Records	14	10	9	8	8	9.8
	MySQL (google.maps.Marker)					
Test Order	**1**	**2**	**3**	**4**	**5**	**Average**
1,000 Records	1052	1039	1041	1049	1048	1045.8
5,000 Records	1128	1138	1143	1111	1116	1127.2
10,000 Records	1202	1194	1230	1233	1240	1219.8
50,000 Records	1842	2145	1915	1867	2211	1996.0
100,000 Records	3050	3332	2938	2895	3123	3067.6
						Unit: Milliseconds (ms)

2.4.3 Limitation and scalability

Fusion Table has some limitations on storage and usage. Each user can import data files no more than 100 MB into each Fusion Table, and each Google cloud account can contain data no more than 250 MB. The Google Fusion Table is an experimental product, which does not have a payment option for increasing the storage space. However, the data inside the Google Fusion Table is text-based which takes up very little space. When data are inserted into the Google Fusion Table, efforts have been made with additional code/scripts to save space by normalizing each field and trimming unnecessary spaces. Currently, there are 2730 records in the Fusion Table, which takes up 657 KB out of 250 MB. This means approximately 1 million similar data records can be stored with

just this one Google cloud account. Furthermore, photo submissions are uploaded to a separate server with terabyte-level shared storage space and only the URLs linked to the photos are stored in Fusion Table.

The situation when the dataset grows beyond the limit of approximately 1 million records has also been taken into consideration. One solution is to have the data stored in multiple Fusion Tables of multiple Google accounts and perform a cross-table query. Another way is to use other cloud-based services, such as Google Cloud SQL and BigQuery, Amazon EC2, and Windows Azure. Google services will be our first choice because it is usually straightforward to develop applications with other Google products, such as Google Maps/Earth and Google Chart.

When inserting a data record into the Fusion Table, the record should be less than 1 MB, and a maximum of 25,000 requests per day can be sent to one Google account with free Fusion Table API access. However, the number of maximum requests per day can be increased by request through Google.

As a result, there is a trade-off between using Fusion Table resources directly and consuming a small portion of the resources from clients. In order to reduce the times in querying the Fusion Table, data from the prior queries are stored on the client side in the global flood cyber-infrastructure. If the next operation from the client side returns the same result as the previous operation, no request will be sent to the Fusion Table. It will use the stored data instead.

2.4.4 Data sharing

Although Google Fusion Table API does not provide a way to download raw data programmatically, as a shared cyber-infrastructure, the data in the Fusion Table of CyberFlood is free to download. A link can be provided to the actual Fusion Table from where users can view raw data and download them as a CSV or KML file. After the raw data have been made accessible, it is possible for users to adapt the raw data to visualize flood events in their own way and gain more discovery.

2.4.5 Sustainability

In order to involve people, some poster presentations about CyberFlood have been given at several conferences. Meanwhile, iOS apps for iPad and iPhone are under development (Figure 2.7), providing functions for people to view flood events visualizations in the form of map and chart and submit their witness accounts of flood events. Plans are made to advertise the CyberFlood through non-traditional media, such as social media Facebook and Twitter. We have also developed the mPING (Meteorological Phenomena Identification Near the Ground: http://www.nssl.noaa.gov/projects/ping/) app which includes flood entries (4 levels of severity) and uses crowdsourcing technique to obtain data. Given that the mPING has more than 200,000 active users today, this app will also be utilized to advertise our CyberFlood system.

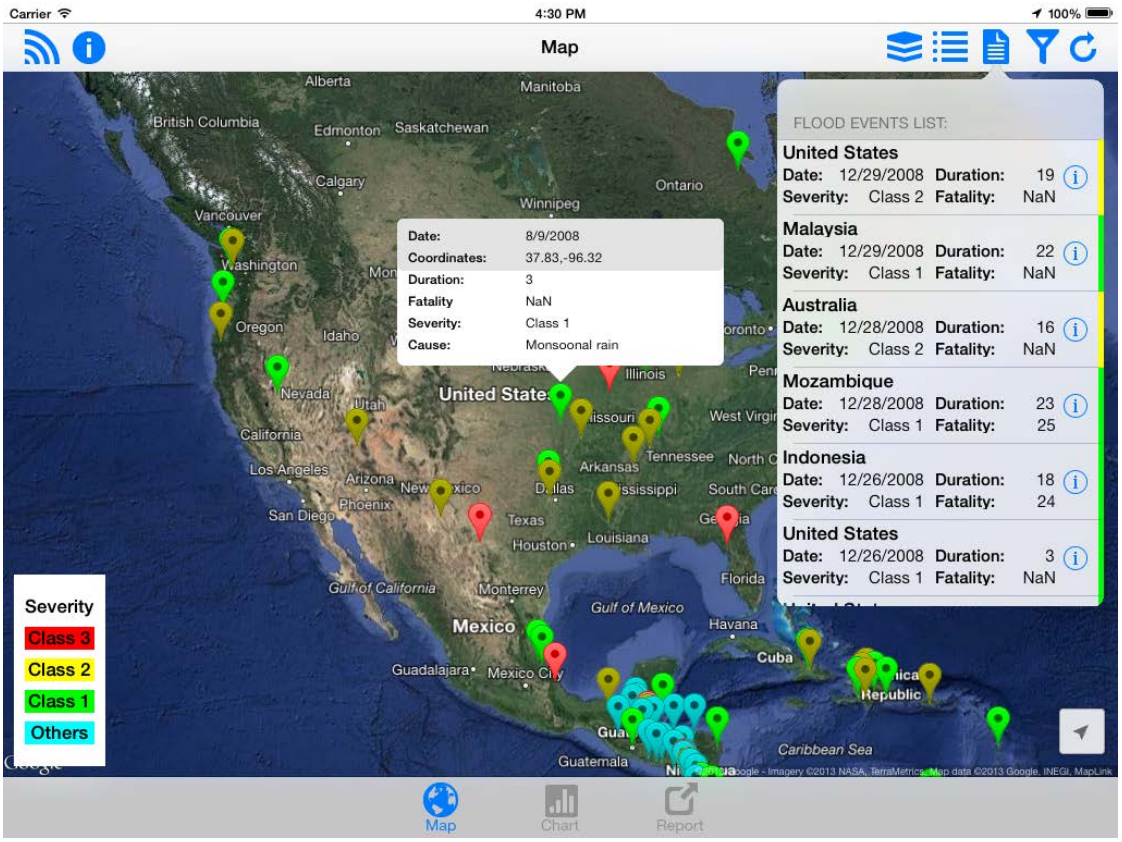


Figure 2.7 Mobile version of the cyber-infrastructure.

Since only limited entries from crowdsourcing during the 2009-2013 period are obtained, locally recruited students are compiling flood events from multiple sources for that period with manual quality control. Data for these years will be available in CyberFlood.

2.5 Conclusion

The cyber-infrastructure for global flood disaster community (CyberFlood), with cloud computing service integration and crowdsourcing data collection, provides on-demand, location-based visualization, as well as statistical analysis and graphing functions. It involves citizen-scientist participation, allowing the public to submit their personal accounts of flood events to help the flood disaster community to archive comprehensive information of flood events, analyze past flood events, and get prepared for future flood events. This cyber-infrastructure presents an opportunity to eventually modernize the existing methods the flood disaster community utilizes to collect, manage, visualize, and analyze data with flood events.

In the future, data describing the flood reports in this cyber-infrastructure will be linked to real-time and archived satellite-based flood inundation areas, observed stream flow, simulated surface runoff from a global distributed hydrological modeling system, and precipitation products. These datasets will be beneficial both as method to validate the crowdsourced flood events and to help educate, motivate, and engage citizen-scientists about the latest advances in satellite remote-sensing and hydrological modeling technologies. Given the elasticity of a cloud-based infrastructure, this cyber-infrastructure for global floods can be applied to other natural hazards, such as droughts and landslides, at both global and regional scales.

Chapter 3: Water balance-based actual evapotranspiration reconstruction from ground and satellite observations over the conterminous United States

The objective of this study is to produce an observationally based monthly evapotranspiration (ET) product using the simple water balance equation across the conterminous United States (CONUS). We adopted the best quality ground- and satellite-based observations of the water budget components, i.e., precipitation, runoff, and water storage change, while ET is computed as the residual. Precipitation data is provided by the bias-corrected PRISM observation-based precipitation dataset, while runoff comes from observed monthly streamflow values at 592 USGS stream gauging stations that have been screened by strict quality controls. We developed a land surface model-based downscaling approach to disaggregate the monthly GRACE equivalent water thickness data to daily, 0.125º values. The derived ET computed as the residual from the water balance equation is evaluated against three sets of existing ET products. The similar spatial patterns and small differences between the reconstructed ET in this study and the other three products show the reliability of the observationally based approach. The new ET product and the disaggregated GRACE data provide a unique, important hydro-meteorological data set that can be used to evaluate the other ET products as a benchmark dataset, assess recent hydrological and climatological changes, and terrestrial water and energy cycle dynamics across the CONUS. These products will also be valuable for studies and applications in drought assessment, water resources management, and climate change evaluation.

3.1 Introduction

As one of the major components of the global hydrologic cycle, evapotranspiration (ET) is a complicated process and composed of evaporation from land surface and water bodies, and transpiration from vegetation to the atmosphere (Allen et al. 1998). Evaporation and transpiration processes occur simultaneously and are difficult to separate (Anderson et al. 2007, Liu et al. 2011, Mallick et al. 2014). Accurately estimating actual ET is of great importance because it is a crucial variable in water resources management, agriculture, and ecology (Khan et al. 2010), and an important process in the fields of hydrology, meteorology and atmospheric sciences (Chauhan and Shrivastava 2009).

Several approaches have been developed to estimate actual ET, including meteorology-driven diagnostic models such as the Penman-Monteith (PM) method (Monteith 1965), satellite data-driven PM approaches (Cleugh et al. 2007, Mu et al. 2007, Zhang et al. 2009, Zhang et al. 2010, Zhang et al. 2008), satellite data-driven Priestly-Taylor empirical approach (Fisher, Tu, and Baldocchi 2008), energy balance methods (Bastiaanssen et al. 1998, Su 2002, Wang and Bras 2011, Wang and Bras 2009), vegetation index-ET empirical relationship methods (Gillies, Kustas, and Humes 1997, Nishida et al. 2003, Tang et al. 2009), and data-driven statistical methods (Jung et al. 2010). The water balance approach is another way to determine ET by quantifying it as the residual in the water balance equation. This method is simple and sound in theory, and warrants accurate estimate of ET as long as the other water components can be accurately measured. Additionally, unlike the other approaches, it does not require additional meteorological inputs except precipitation. One good example for

measuring/estimating ET using the water balance approach is the lysimeter. The water balance method has been used to estimate ET in previous studies (Long, Longuevergne, and Scanlon 2014, Ramillien et al. 2006, Zeng et al. 2012, Zhang et al. 2010), but this approach is usually applied to one or multiple basins to derive the areal-mean ET of these basins that serve as an ET validation data set.

The recent ET estimates by model simulations and satellite-driven algorithms are usually evaluated against point FLUXNET eddy covariance measurements (Mu et al. 2007, Velpuri et al. 2013, Zhang et al. 2009) and simulations from land surface models (Jung et al. 2010, Schwalm et al. 2013). Few of these studies use basin-wide ET estimates from water balance computations as benchmark values to evaluate the remotely sensed ET estimates (Zeng et al. 2012, Zhang et al. 2010). The water balance-based ET is rarely available, covers few regions, and has coarse spatial resolution due to the limited data availability and continuity.

To produce a subbasin-wide ET product with continuous temporal coverage and downscaled gridded water storage change data with a relatively finer spatial resolution (0.125°), we utilized the trustworthy ground- and satellite-observed hydrological data provided by USGS, NASA, and USDA to estimate monthly actual ET and monthly 0.125° water storage change data from April 2002 to September 2013 across the conterminous United States (CONUS). The method developed in this study computes actual ET as the residual in the simple water balance equation. The objective of this study is to produce an observationally based monthly evapotranspiration (ET) product using the simple water balance equation across the CONUS. This dataset can be used to evaluate the other ET products as a benchmark dataset, assess recent hydrological and

climatological changes across the CONUS. These products will be also valuable for studies and applications in drought assessment, water resources management, and climate change evaluation.

3.2 Data and Methodology

3.2.1 Study Area and Data

The spatial domain of this study is the CONUS, ranging from 25°N to 50 °N and from 124.75 °W to 67 °W (Figure 3.1). The data used in this study include observations of precipitation, runoff, and water storage change from ground and satellite data, and river network and topographical data from a remote sensing-derived digital elevation model (DEM). The river network data have a spatial resolution of 0.125° and were derived from an upscaled global data set from the combined HydroSHEDS and HYDRO1K datasets (Wu et al. 2012).

Figure 3.1 Locations of 592 USGS stream gauging stations used in this study and spatial distributions of their corresponding sub-basins over the CONUS; the blank areas are regions without sufficient good-quality observational data.

The precipitation data are from the PRISM (Parameter-elevation Regressions on Independent Slopes Model) daily precipitation data set produced by the PRISM group at Oregon State University (http://www.prism.oregonstate.edu). The PRISM daily precipitation product is a 4-km gridded estimate of precipitation for the CONUS based on observations from a wide range of monitoring networks with sophisticated quality control, and bias and topography corrections (Daly et al. 2008). The PRISM interpolation method calculates climate–elevation regression for each grid cell, and stations entering the regression are assigned weights based primarily on the physiographic similarity of the station to the grid cell. Factors considered are location, elevation, coastal proximity, topographic facet orientation, vertical atmospheric layer, topographic position, and orographic enhancement caused by the underlying terrain (Daly et al. 2008). The PRISM data set is the source of USDA's official climatological data. In this study, all analyses were conducted on a geographical grid with a resolution of 0.125°. Therefore, the PRISM precipitation was first aggregated from 4 km to 0.125° and then summed from daily values to monthly values.

Monthly mean streamflow observations from all USGS stream gauging stations, which have continuous discharge data between April 2002 and September 2013, were chosen to derive the monthly runoff depth at the sub-basin level. Some of these stations were further screened out if differences between their drainage areas as provided by USGS metadata and the areas derived from the 0.125° DEM-based flow accumulation are larger than 20%. If multiple streamflow measurement stations fall in the same 0.125° grid cell, only the station with the largest drainage area was kept for further analysis. The drainage area of each station must contain at least two 0.125° grid cells. After the strict

screening process, streamflow data from 592 USGS stations were chosen for further analysis (Figure 3.1).

Monthly equivalent water thickness (EWT) of water storage is provided by the Gravity Recovery and Climate Experiment (GRACE) satellite-derived data set. GRACE is a twin-satellite mission launched in March 2002 to observe the variation of Earth's gravity field anomalies. GRACE satellites provide information on changes in the gravity fields, which are controlled primarily by variations in water distribution and are used to derive terrestrial water storage change at a spatial resolution of ≥200,000 km^2 (Tapley, Bettadpur, Ries, et al. 2004). The latest GRACE land grid data Release-05 (RL05) released in February 2014 is used in this study. The RL05 is a level-3 GRACE product containing the EWT product in centimeters with a spatial resolution of 1°×1° (Chambers 2006). This gridded data set was converted from sets of spherical harmonic coefficients of the standard GRACE product describing the monthly variations in Earth's gravity filed after applying a series of GRACE filters (Swenson and Wahr 2006, Wahr, Molenaar, and Bryan 1998, Wahr, Swenson, and Velicogna 2006). Gridded scaling factors are also applied to the gridded GRACE EWT to minimize the leakage error due to resampling and post-processing, i.e., the filtering and smoothing processes (Landerer and Swenson 2012). Although GRACE provides an opportunity to better constrain the water budget equation, it has relatively coarse spatial resolution and suffers periodic data gaps due to battery management issues and during certain orbit periods (http://grace.jpl.nasa.gov/data/gracemonthlymassgridsoverview/). To achieve a continuous terrestrial water storage change data with a spatial resolution of 0.125°, we developed a downscaling approach in which the GRACE data were used to constrain the

water storage thickness simulated by four land surface models (LSMs) from North American Land Data Assimilation System project phase 2 (NLDAS-2) and correct the bias in the modeled water storage thicknesses. Details of this downscaling method are described in Section 3.2.2.

3.2.2 Methodology

In this study, we derived monthly areal-mean actual ET on a sub-basin level using the water balance equation by assuming no net groundwater flow across the boundary of a river basin of interest:

$$ET = P - R - \Delta S + \varepsilon \tag{3.1}$$

where P (mm) is the monthly precipitation; R (mm) is the monthly runoff depth; ΔS (mm) is the monthly terrestrial water storage change, i.e. change in the monthly EWT; and ε is an error term. Because the water budget terms (P, R and ΔS) are derived from ground and satellite observations, there are some measurement and processing errors in these data sets (Daly et al. 2008, Landerer and Swenson 2012, Swenson and Wahr 2006, Tapley, Bettadpur, Watkins, et al. 2004). However, quantifying the error for each of the datasets for each sub-basin is impractical, and we assume that the errors are random and small in magnitude relative to the values of the water balance variables. Therefore, the derived monthly ET values inherit these errors given that they are computed as the residual. The sources and detailed processing of the three water budget terms used to compute the ET are described in Section 3.2.1 and the remaining part of this section.

Calculation of sub-basin runoff depth

Since many of these USGS streamflow measurement stations are nested within the same parent watersheds (Figure 3.1), we first derived the topological relationships among these stations within the same parent basins from the river network data (i.e., flow direction, flow accumulation area). The drainage areas of all neighboring upstream stations from a given station were subtracted out from the drainage area of this station so that each station was attributed to unique contributing areas, i.e. a sub-basin associated to a specific station does not contain or overlap with other sub-basins. For example, there are 102 stations in the Missouri river basin; therefore, the application of the above procedure produces 102 sub-basins that do not overlap with each other (Figure 3.1).

Missing values exist in some of the 592 USGS stations for different reasons, but these data gaps must be less than 20% of the total record, else they are removed. Linear interpolation is not a good solution when the data gap encompasses two or more months. This is because linear interpolation can artificially smooth the fluctuation of monthly discharge values. Instead, we applied an alternative method in which the multi-year mean value of a missing month ($\overline{Q_m}$), the discharge of its nearest month (Q_n), and the multi-year mean value of the nearest month ($\overline{Q_n}$) are used to fill the missing value of the missing month (Q_m):

$$Q_m = \frac{\overline{Q_m} \times Q_n}{\overline{Q_n}} \tag{3.2}$$

In essence, we assume that the ratio of monthly discharge in a missing month to its multi-year mean is equal to the ratio of monthly discharge in its nearest month to the multi-year mean discharge of the nearest month.

The monthly runoff depth of sub-basin i is then computed by the following equation:

$$R_i = \frac{(Q_i - \sum_{n=1}^{N} Q_n) \times T}{A_i} \times 1000 \quad (3.3)$$

where R_i is the monthly runoff depth of sub-basin i (mm); Q_i is the monthly discharge at station i ($m^3\ s^{-1}$); Q_n is the monthly discharge of neighbor upstream station n of station i ($m^3\ s^{-1}$); N is the total number of neighbor upstream stations for station i; A_i is the contributing land area of sub-basin i (m^2); and T is time (s) in a month.

Downscaling of GRACE equivalent water thickness data

As we discussed previously, the GRACE data have periodic gaps and a coarse spatial resolution. To utilize the GRACE data to derive continuous, finer resolution time series of water storage change, we developed a model-based approach to downscale the GRACE data. First, hourly 0.125° simulations of the Variable Infiltration Capacity (VIC) , Noah Land Surface Model (Noah), Mosaic, and Sacramento Soil Moisture Accounting (SAC) models from North American Land Data Assimilation System project phase 2 (NLDAS-2) were used in this study to estimate daily water thickness of soil water storage across the CONUS. The four models form the land surface model (LSM) ensemble executed over the CONUS in NLDAS-2 (Xia, Mitchell, Ek, Sheffield, et al. 2012). The VIC model is a semi-distributed grid-based land surface hydrological model, which solves for full water and energy balances (Liang et al. 1994, Liang, Wood, and Lettenmaier 1996). The Noah model is a community LSM, which simulates soil moisture (both liquid and frozen), soil temperature, skin temperature, snowpack depth, snowpack water equivalent (and hence snowpack density), canopy water content, and the energy

flux and water flux terms of the surface energy balance and surface water balance (Chen et al. 1996, Ek et al. 2003, Koren et al. 1999, Mitchell et al. 2004). The Mosaic model was developed for use in NASA's global climate model and simulates energy and energy balance, and soil moisture and temperature (Koster, Suarez, and Heiser 2000). Originally formulated as a lumped conceptual hydrological model, SAC has since been converted into a distributed version and has adopted some components of the surface-vegetation-atmosphere transfer scheme developed within the coupled climate modeling community (Koren et al. 2007). In the NLDAS-2 project, the VIC model is equipped with three soil layers with a fixed 10 cm top layer and two other layers with spatially varying thicknesses, while the Noah model has spatially uniform four soil layers with fixed thicknesses of 10, 30, 60, and 100 cm (Xia, Mitchell, Ek, Sheffield, et al. 2012). Mosaic has three soil layers with thicknesses of 10, 30, and 160 cm, while SAC has five soil layers to cover a 2-m soil profile (Xia, Mitchell, Ek, Sheffield, et al. 2012).

To downscale the GRACE data, we first aggregated four sets of hourly LSM data separately to produce four sets of daily equivalent water thickness (EWT: mm) data on a 0.125° grid. The EWT is the integral of water above and inside the soil column within each grid cell, including surface water and soil water computed in the four LSMs. However, like many other LSMs, the four NLDAS-2 LSMs do not simulate groundwater fluxes (Xia et al. 2015, Xia, Mitchell, Ek, Cosgrove, et al. 2012, Xia, Mitchell, Ek, Sheffield, et al. 2012); thus, the models do not account for changes in groundwater fluxes such as water depletion and recharge. However, these changes can be captured by the GRACE data over large spatial extent. We then normalized the daily EWT by its mean value from January 2004 to December 2009 grid cell by grid cell to produce normalized

EWT (S_i) by following the same normalization procedure used in the GRACE data (http://grace.jpl.nasa.gov/data/gracemonthlymassgridsoverview/). Considering that the footprint of GRACE signals is ~200,000 km^2 (about 4° by 4°) (Longuevergne, Scanlon, and Wilson 2010) and the GRACE data are believed to have large uncertainty for resolutions < its footprint (Long, Longuevergne, and Scanlon 2014, Longuevergne, Scanlon, and Wilson 2010), we aggregated the 1° GRACE data to 4° and then downscaled the 4° data to 0.125° using the following method. The 0.125° normalized EWTs from the LSMs were aggregated to 4.0° to match with the 4°GRACE grid, using area as a weighting factor as:

$$S_M = \frac{\sum(S_i \times a_i)}{\sum a_i} = \frac{\sum(S_i \times a_i)}{A} \tag{3.4}$$

where S_M (mm) is the 4.0° LSM normalized EWT; a_i (m^2) is the area of the 0.125° grid cell *i*; A (m^2) is the total area of the 4.0° grid cell containing the 0.125° grid cell *i*. The difference between the 4.0° LSM normalized EWT and 4.0° GRACE normalized EWT (S_G) represents the bias (B) of the modeled EWT if we treat the GRACE data as "truth":

$$B = S_M - S_G \tag{3.5}$$

The total water volume offset ($B \times A$) between the model and GRACE data were further distributed to the 0.125° grid using water volume as weight:

$$b_i = \frac{B \times A \times \frac{S_{oi} \times a_i}{\sum(S_{oi} \times a_i)}}{a_i} = \frac{B \times A \times S_{oi}}{\sum(S_{oi} \times a_i)} \tag{3.6}$$

where b_i is the bias of the 0.125° model EWT; and S_{0i} is the pre-normalized 0.125° model EWT. Since the GRACE data is a monthly composite product and different number of daily measurements is used for different months to calculate monthly values,

the bias b_i is treated as the bias in the middle of a month. Then linear interpolation is applied to produce daily bias values for each grid cell. Finally, once the bias b_i is subtracted from S_i, we can obtain the 0.125° bias-corrected daily EWT (S_i'):

$$S_i' = S_i - b_i \tag{3.7}$$

This downscaling method preserves the accuracy of the GRACE data and provides that the summation of the 0.125° bias-corrected EWT over any 4° GRACE grid cell is equal to the original 4° GRACE value at the same grid cell. Moreover, this downscaling method produces a finer resolution, continuous daily EWT series. The monthly water storage change (ΔS_m) in month *m* at grid cell *i* is derived as the difference between bias-corrected daily EWT value on the last day of a given month and on the last day of its previous month as:

$$\Delta S_m = S_i'(d_m) - S_i'(d_{m-1}) \tag{3.8}$$

where d_m and $d_{m-1}$are the Julian days of months *m* and *m-1*, respectively.

Since we downscaled the GRACE data using the outputs from four LSMs, we correspondingly produced four sets of 0.125° ΔS_m and monthly actual ET. The four sets of data form an ensemble. We used the ensemble mean as the final product. Hereafter, the reconstructed ET and downscaled ΔS_m denotes their ensemble means except as otherwise noted. To quantify the uncertainty in the reconstructed ET due to difference in the model outputs, we applied the commonly used ensemble standard deviation (SD), i.e. ensemble spread, as a metric:

$$SD = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(ET_m - \overline{ET})^2} \tag{3.9}$$

where

$$\overline{ET} = \frac{1}{M}\sum_{m=1}^{M} ET_m \tag{3.10}$$

and M (=4) is the number of ensemble members. Considering that there is only one set of precipitation and runoff data, the ensemble spread of ΔS is essentially the same as that of ET according to equation (3.1).

Evaluation of the water balance based ET

To evaluate the reconstructed ET values using the subbasin water balance approach, we compared the ET estimates with three data sets of ET estimations with reported good quality. One ET data set is produced by a remote sensing driven process-based algorithm (Zhang et al. 2010), the second data set is a data-driven, upscaled eddy-covariance flux measurements from the global FLUXNET work using a sophisticated machine learning method (Jung et al. 2010), and the third data set is the MOD16A2 global ET product (Mu, Zhao, and Running 2011). All of the three ET data sets are widely assessed and used in the atmospheric and earth sciences community (Cai et al. 2011, Wang and Alimohammadi 2012, Zeng et al. 2012), and are treated as benchmark ET products in some studies (Swenson and Wahr 2006, Zeng et al. 2012).

Three statistical variables were used to measure the similarity between the three products, including mean difference (MD), root mean square difference (RMSD) and the coefficient of determination (R^2). The mean difference is defined as the average difference between the estimates to be evaluated (y_i) and the estimates to be compared against (x_i):

$$\mathrm{MD} = \frac{\sum_{i=1}^{n}(y_i - x_i)}{n} \tag{3.11}$$

where n is the sample size. RMSD is used to measure the closeness between two ET products and defined as:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}} \tag{3.12}$$

The R^2 coefficient is used to evaluate the covariance between the two estimates of ET.

3.3. Results

3.3.1 Downscaled EWT and its Spatiotemporal Patterns

Figure 3.2 shows the normalized regional mean EWT values over the CONUS and its twelve hydrologic regions from Apr. 2002 – Sep. 2013 using the original monthly GRACE data, the original model daily EWT, and the downscaled daily EWT. Although there are some discrepancies between the GRACE data and the original ensemble mean of EWT from the NLDAS-2 LSMs, the mean of model results shows a generally good agreement with the GRACE data in terms of the seasonality and interannual variability (Figure 3.2). It is clear that the downscaled daily EWT matches the original GRACE data quite well with the added benefit of improved resolution using the model-based downscaling technique (Figure 3.2). The EWT series shows a clear, consistent seasonality with peak values falling between February and April when snow storage reaches maximum values and with minimum values around September when air temperatures are high accompanied by low seasonal precipitation in most hydrological units of the CONUS (Figure 3.2). It also shows large inter-annual variability; the difference between the highest water storage and the lowest water storage during the twelve years is about 180 mm, which is equivalent to 1,055 km^3 of liquid water. The min-max spreads of the original model water storage and the downscaled GRACE data (grey areas in Figure 3.2) are generally narrow with relatively large spreads in few months in a couple of hydrological regions, e.g. the Northwest and Southeast regions (Figure 3.2), indicating that the difference between the NLDAS-2 LSMs water storage data are subtle in these large regions.

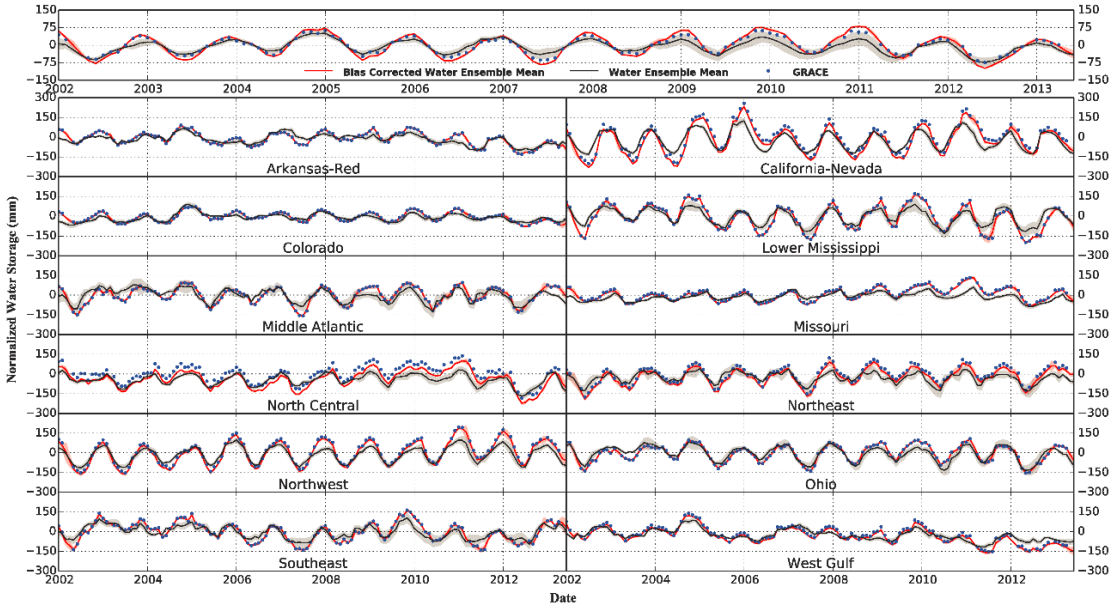


Figure 3.2 Time series of monthly terrestrial water storage change over CONUS and its twelve hydrologic regions from the original and land surface model-based downscaled GRACE data from 2002 to 2013; the downscaled data are the ensemble mean, while the grey area denotes the min-max ensemble range.

The spatial pattern of the eleven-year (Apr 2002-Mar 2013) mean water storage change shows that most of the CONUS had very small water storage changes (Figure 3.3(a)), indicating most of these areas are in a water storage balanced state. However, some areas in the southern CONUS (e.g. eastern Texas and western Louisiana) and central Minnesota show negative multi-year water storage change, implying that these areas have lost water in the past twelve years. The loss of water storage in these areas is largely attributed to groundwater depletion and recent drought episode (Freshwater Society 2013, Long et al. 2013). In contrast, part of Florida shows a small gain of water storage during the past eleven years (Figure 3.3(a)).

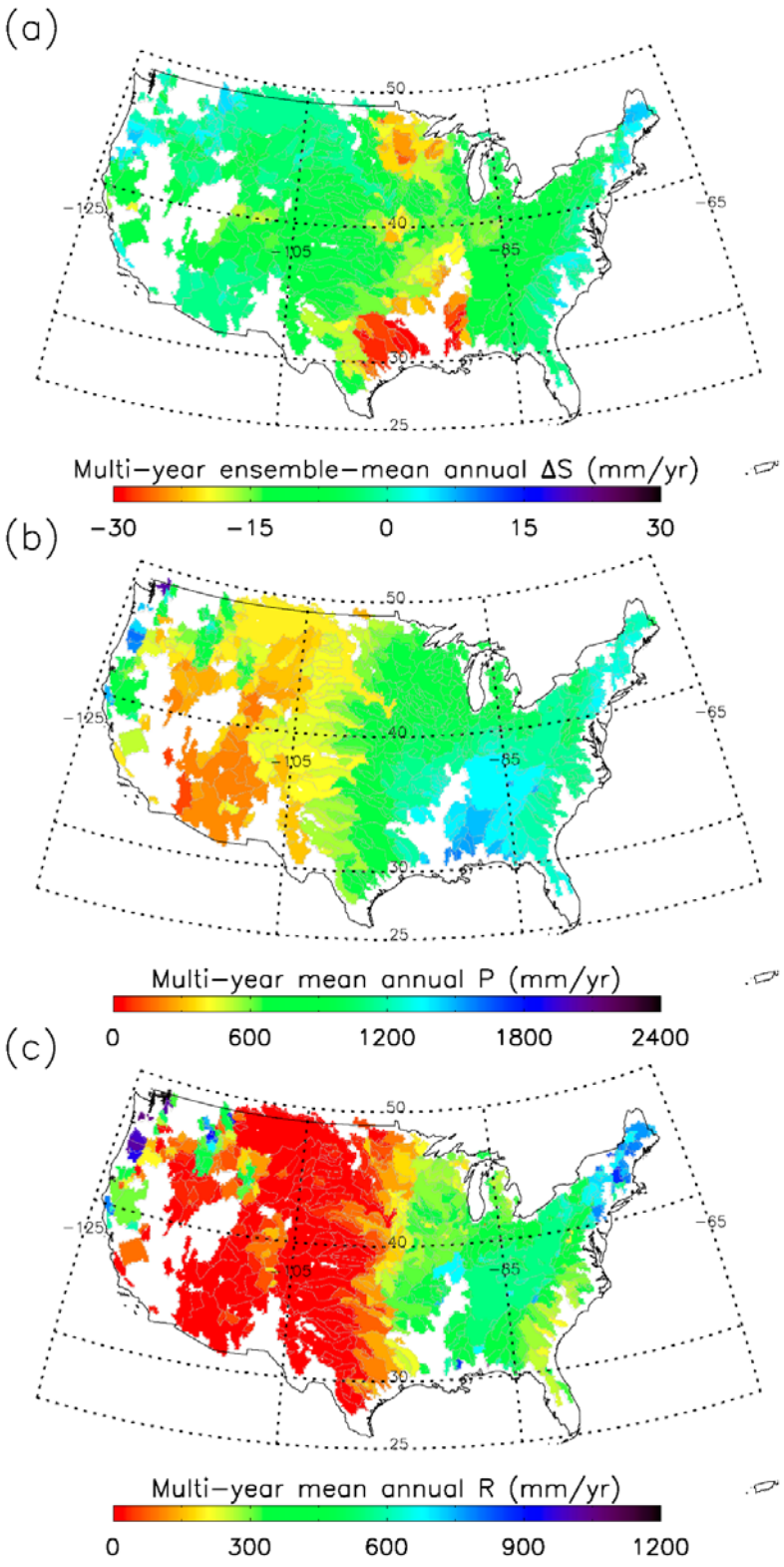


Figure 3.3 Spatial patterns of ground and satellite observed multi-year (from Apr 2002 to Mar 2013) mean annual (a) ensemble-mean terrestrial water storage change (ΔS), (b) precipitation (P), and (c) runoff depth (R).

3.3.2 Spatial Patterns of Water Budget Terms in CONUS

Spatial patterns of ground-observed, twelve-year mean annual precipitation and annual runoff depth are shown in Figure 3.3. The mean annual precipitation displays a clear spatial gradient in which annual precipitation gradually decreases from the Southeast US to the Midwest and to the Rocky Mountains, and then increases from the Rocky Mountains to West Coast (Figure 3.3(b)). The spatial pattern of runoff depth is very similar to that of precipitation with a correlation coefficient of 0.84 (P<0.001); the west and east coasts of the US and the Southeast have the highest annual runoff, while the Rocky Mountains and the Great Plains have the lowest annual runoff (Figure 3.3(b)). The similarity between the spatial patterns of precipitation and runoff indicates that precipitation is the major controlling factor of runoff.

3.3.3 Evaluation of Water Balance-based ET Reconstruction and its Spatial Pattern

Multi-year average annual ET from the ensemble mean of water balance-based reconstructions (ET_{Recon}; Figure 3.4(a)) is compared with the remote sensing-based estimate (Zhang et al. 2010) (ET_{Zhang}; Figure 3.4(b)), the data-driven upscaled estimate (Jung et al. 2010) (ET_{Jung}; Figure 3.4(c)), and the MOD16 ET product (ET_{Mu}; Figure 3.4(d)). ET estimates from all four methods show similar spatial patterns. ET is the highest in the Southeast and decreases westward and northward, and reaches its minimum in the interior of the Intermountain West such as the deserts in Nevada. ET increases again from the Intermountain West to the West coast (Figure 3.4). Although ET_{Recon} has a similar pattern as those of precipitation and runoff, the correlation coefficient of ET_{Recon}

and precipitation is 0.72 ($P < 0.001$); i.e., weaker than that of runoff and precipitation. This is because ET is not only largely controlled by precipitation but also impacted by other factors such as land-cover type, radiation, humidity, wind speed, temperature, etc.

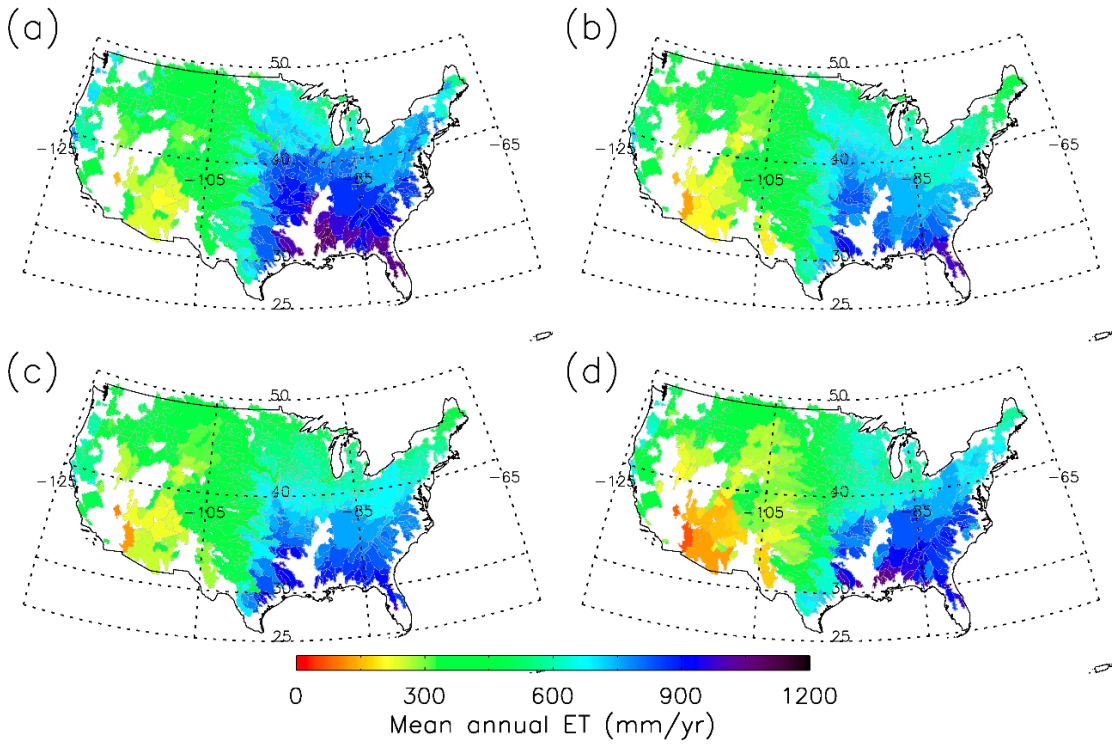


Figure 3.4 Spatial patterns of multi-year average annual ET from (a) the ensemble mean of water balance based reconstructions, (b) a remote sensing based estimate (Zhang et al. 2010), (c) the data-driven upscaled estimate (Jung et al. 2010), and (d) the MOD16A2 product (Mu, Zhao, and Running 2011).

The uncertainty in the reconstructed ET resulted from the difference in the four LSMs outputs is generally small (Figure 3.5): the mean ensemble spread of the reconstructed ET is less than 9 mm/month for 79% of the study region, and the largest ensemble spread is less than 30 mm/month. The regions with relatively large ET ensemble spread are mainly located in the coastal areas and part of the Midwest, while the other regions have generally small uncertainty spread, indicating that the four LSMs have generally compatible spatial patterns of water storage (Figure 3.5).

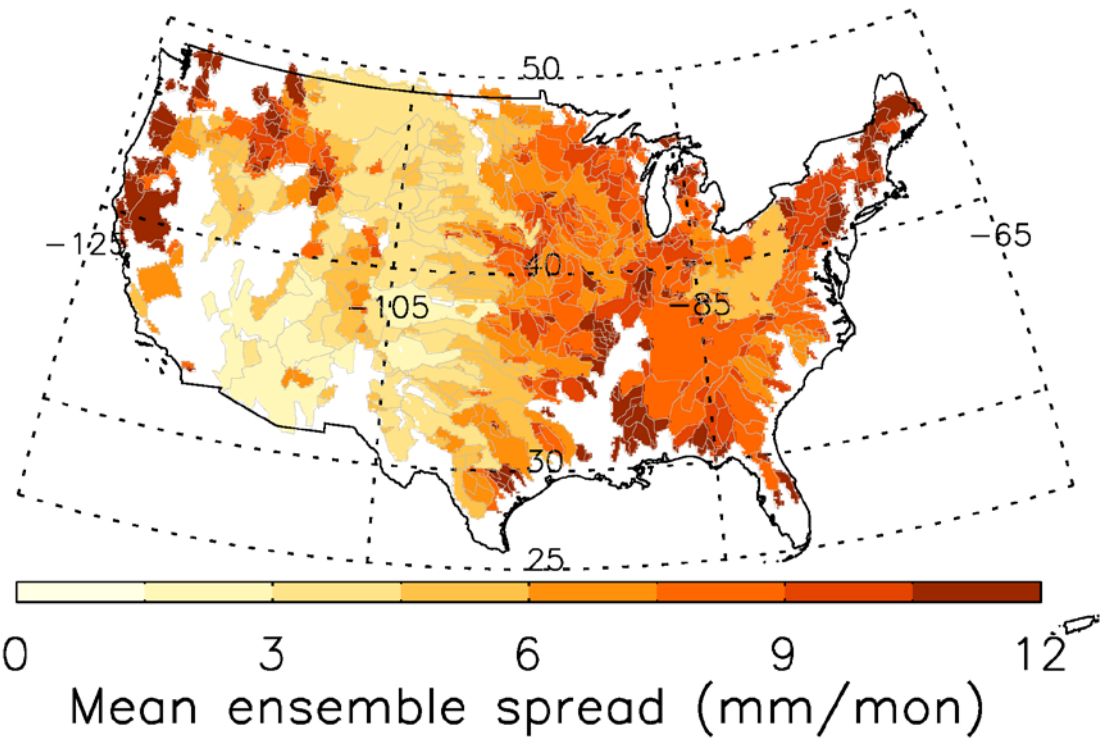


Figure 3.5 Mean ensemble spread of the reconstructed monthly ET.

The four sets of ET estimates across the 592 CONUS sub-basins show very similar spatial gradients (Figure 3.4), although some differences can be noticed. For example, the ET_{Recon} in this study generally has higher values than the other three products in the Southwest (Figure 3.4). The inter-comparison between the four ET estimates show high correlations indicated by the high R^2 values (≥0.74). The mean difference between these ET estimates for the 592 basins ranges from 6.8 to 96.5 mm yr^{-1} (Figure 3.6). The RMSD between the four ET estimates varies between 64.4 and 146.3 mm yr^{-1} (Figure 3.6). It is notable that the ET_{Recon} values show higher similarity and correlation with ET_{Zhang} and ET_{Jung} relative to ET_{Mu} (Figure 3.6a, b, c). In addition, the ET_{Zhang} and ET_{Jung} values are very close to each other and show similar quality. Although the two prior estimates were produced by different approaches, they used similar climatology and remote-sensing data (Jung et al. 2010, Zhang et al. 2010). This may

explain why these two products have very similar results across the CONUS. The generally close spatial patterns and small differences between the four ET estimates from different approaches indicate the high accuracy and robustness of these ET estimates. In other words, the water balance-based ET reconstruction conducted in this study is valid.

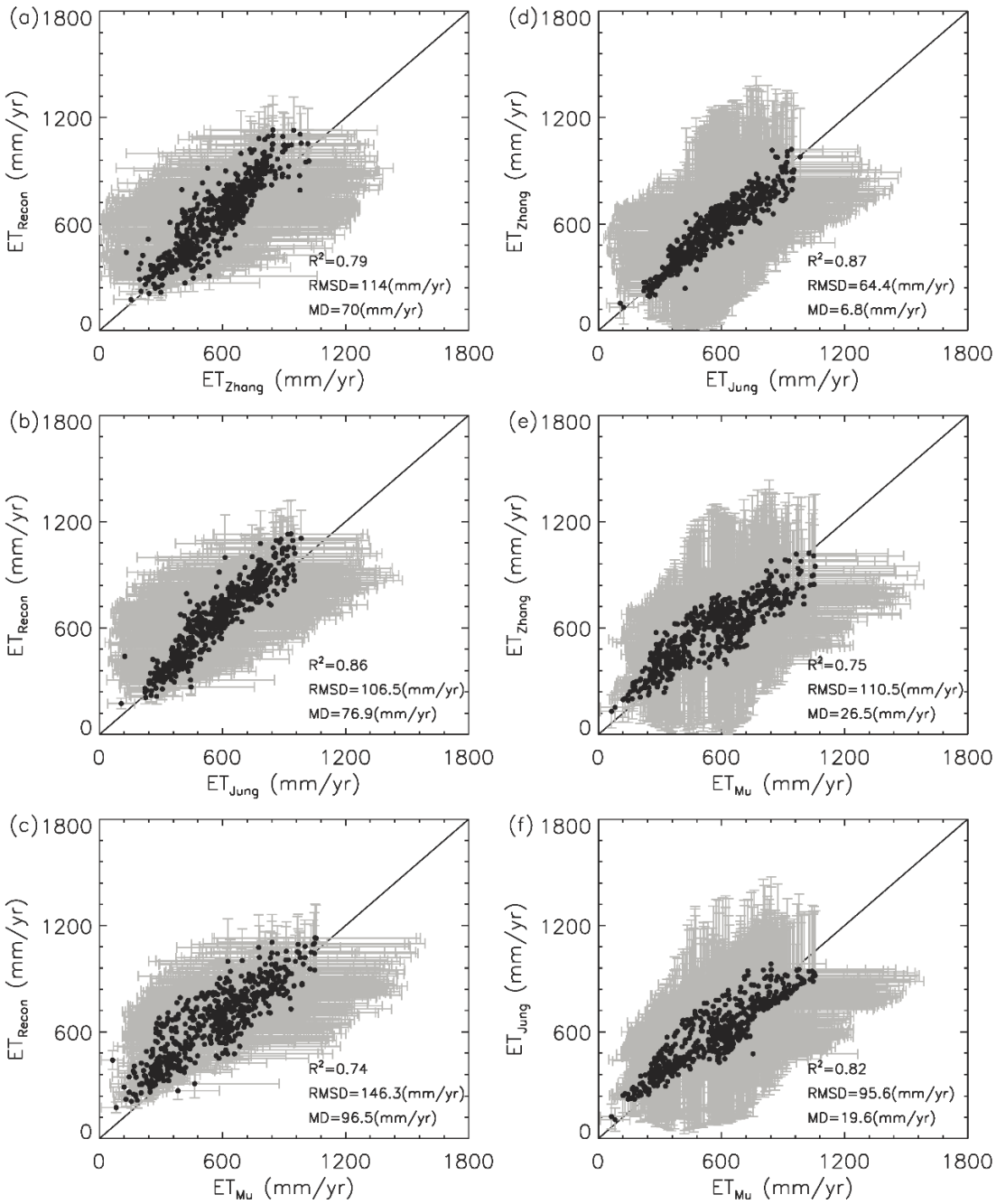


Figure 3.6 Inter-comparisons (a) between mean annual ET estimates from the ensemble mean of water balance based reconstruction (ETRecon) and the remote sensing based estimate by Zhang et al. (2010) (ETZhang), (b) between ETRecon and the data-driven upscaled ET estimate by Jung et al. (2010) (ETJung) , (c) between ETRecon and the MOD16A2 ET by Mu, Zhao, and Running (2011) (ETMu), (d) between ETZhang and ETJung , (e) between ETZhang and ETMu, and (f) between ETJung and ETMu across 592 CONUS basins; black solid circles are basin-level mean annual ET, while grey error bars denotes interannual variability (standard deviation) of basin-level annual ET.

To assess the effectiveness of our downscaling method and the importance of monthly terrestrial water storage change, i.e. the ΔS term, in the water balance based ET estimate, we produced two additional sets of monthly ET records: one is the water balance based ET reconstruction by resampling the 1° GRACE data onto the 0.125° grid using the nearest neighbor method ($ET_{Resample}$), while the other is the ET estimate as the difference between P and R (ET_{P-R}). It is clear that the $ET_{Resample}$ shows substantially poorer agreements with the three independent ET records than the ET_{Recon} in terms of the scatterplots and the R^2 and RMSD metrics (Figure 3.7a-c). This suggests that using the 1° GRACE data without downscaling it to derive sub-basin level ET, in particular for regions that are less than 1° by 1°, will result in additional uncertainty and erroneously abnormal results as shown in Figure 3.7a-c. In other words, our downscaling method has effectively disaggregated the coarser GRACE data to finer (0.125°) resolution, resulting in good-quality ET reconstruction. ET_{P-R} also shows degraded agreements with the three ET records similar to $ET_{Resample}$ in terms of the R^2 and RMSD metrics (Figure 3.7d-f). Like the results of $ET_{Resample}$, the derived ET by ignoring the ΔS term can also result in erroneous and abnormal values such as negative values and erroneously high values as shown in Figure 3.7d-f. Therefore, it is important to account for the ΔS term in order to provide accurate monthly ET estimates using the water balance approach. The downscaling approach implemented in this study is capable of disaggregating the coarser GRACE EWT to finer resolution to achieve reasonably good estimates of ΔS for sub-basins that are even smaller than the footprint of the GRACE data. It is worthy to note that ET_{Recon}, $ET_{Resample}$, and ET_{P-R} all have generally higher values than the three independent remote sensing based ET products, suggesting that these remote sensing

based ET products may tend to understate the actual ET considering these products don't explicitly account for water balance closure and the effect of P on ET.

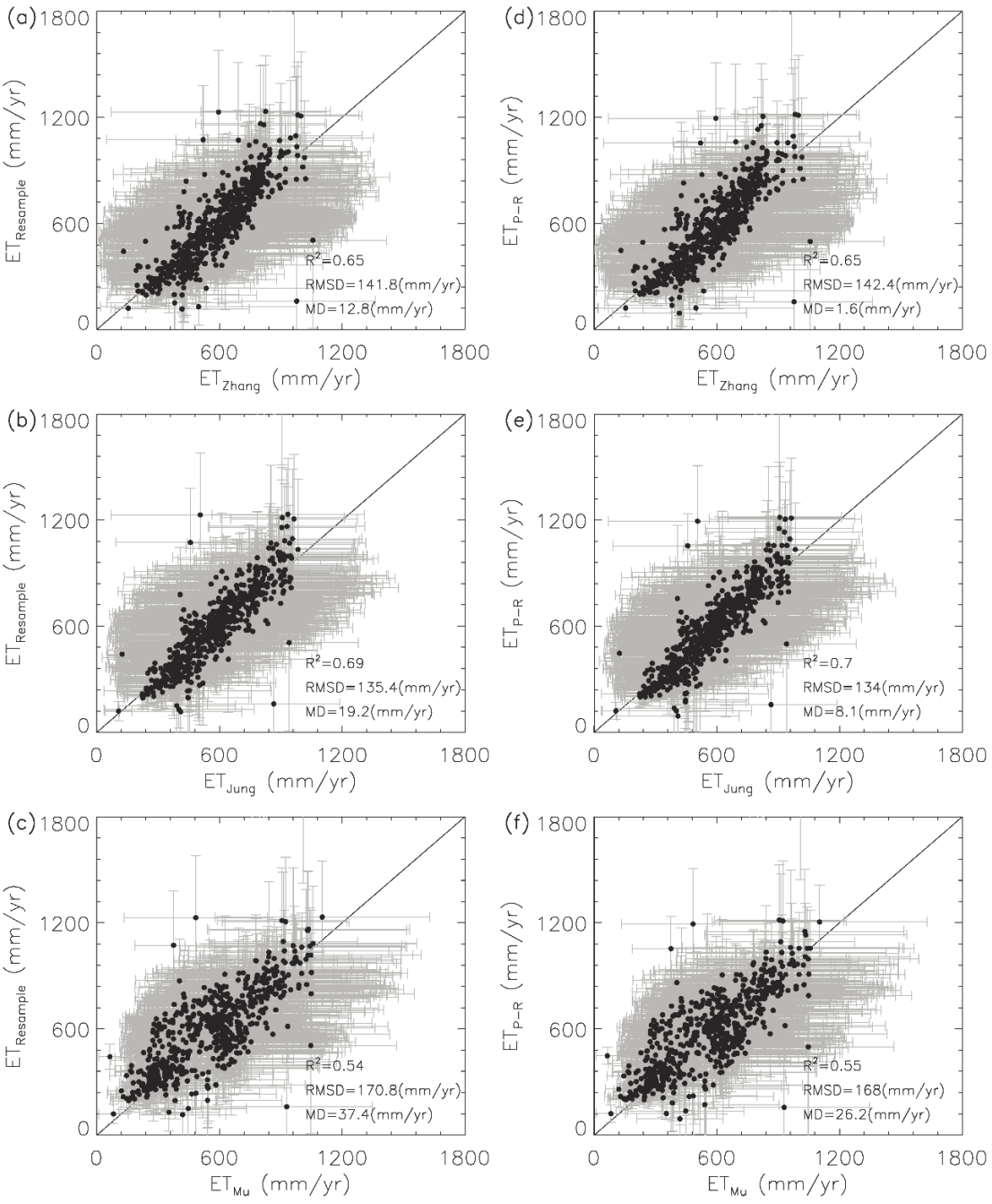


Figure 3.7 Same as Figure 3.6, but for (a-c) intercomparison between the water balance based ET reconstruction by resampling the 1° GRACE data onto the 0.125° grid ($ET_{Resample}$) and the three independent ET records, and (d-f) intercomparison between the ET reconstruction by ignoring change in water storage (ET_{P-R}) and the three ET records.

We further investigated the agreement of seasonality among the four ET estimates, the ET_{Recon} and the three independent ET records, by comparing their twelve-year mean monthly profiles. The results show that all the four ET estimates show similar monthly profiles with peak values in July when solar radiation, temperature and plant growth reach their peaks and with minima in January when solar radiation and temperature reach their minima and most plants are dormant in the CONUS (Figure 3.8). Despite these similar monthly profiles, there are some noticeable differences. For example, the ET_{Recon} has generally higher values than the other three products, especially in the summer months. These differences imply that the existing three ET products may tend to underestimate the actual ET, because the existing ET products do not explicitly quantify some hydrological processes during the frozen periods such as sublimation and snowmelt that impact the ET, and the existing ET products can be also affected by satellite signal saturation during the peak of growing season. It is also notable that ET_{Mu} tends to have lower seasonal variability than the other products indicated by its higher minimum values and smaller peak values. In the rest of the months, ET_{Recon}, ET_{Zhang} and ET_{Jung} products have similar values, while ET_{Mu} have generally lower values than the other products (Figure 3.8).

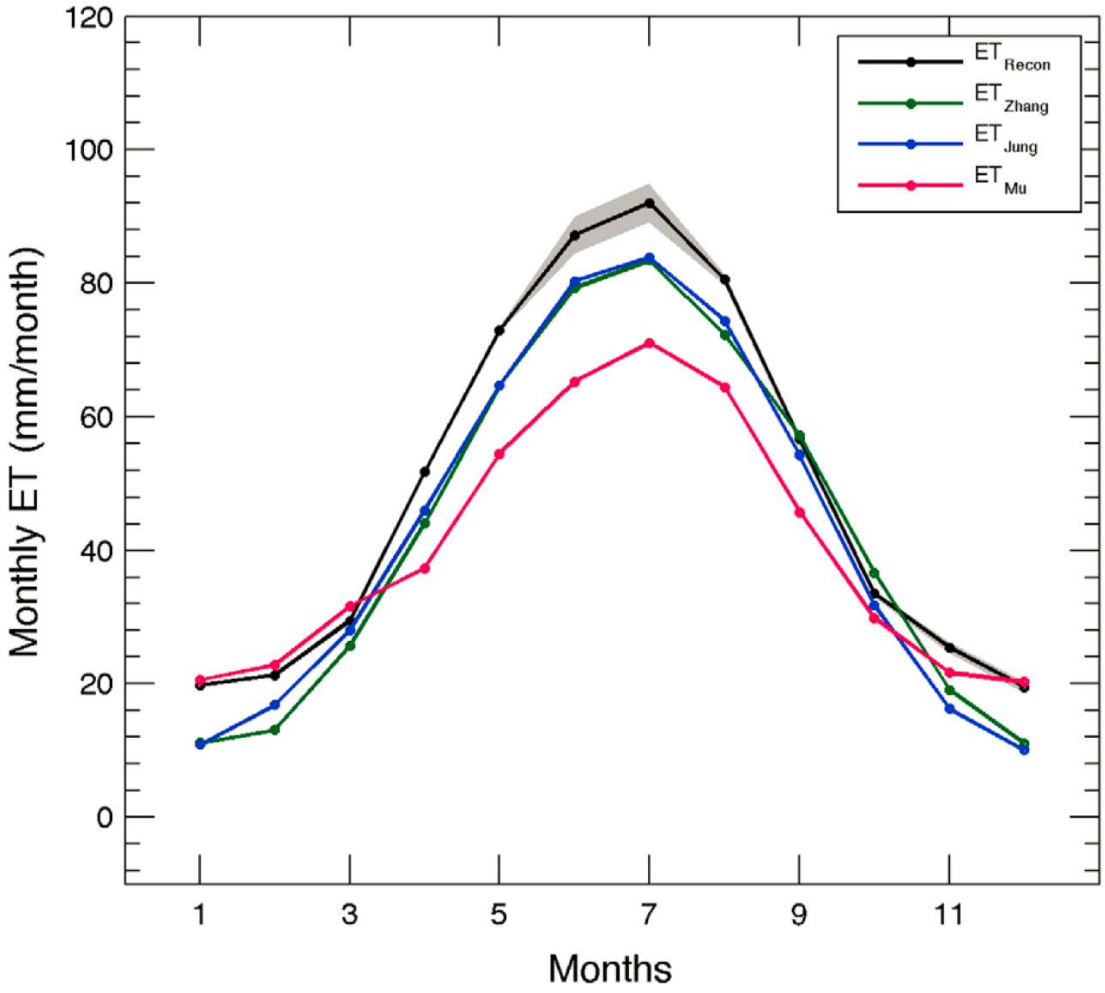


Figure 3.8 Comparison of mean monthly profile of actual ET from the ensemble mean of water balance based reconstructions, remote sensing based estimate (Zhang et al. 2010), data-driven upscaled estimate (Jung et al. 2010) and MOD16A product (Mu, Zhao, and Running 2011).

3.4. Conclusion and Discussion

In this study, a new actual ET product across the CONUS has been derived from high quality satellite and ground observations, including the PRISM precipitation data, USGS observed streamflow data, and GRACE water storage data that has been downscaled using land surface models. This data set covers 73% of the CONUS and is available from Apr. 2002 to Sep. 2013. To our knowledge, this is the first study that derives decadal, continuous monthly ET values across the CONUS from observations using the subbasin water balance method. The method is unique in that it is observationally driven so that ET is computed as the residual in the water balance equation. This differs from past methods and models that often estimate ET using approximate methods and then compute the storage term as the residual in the water balance. The wide availability and accuracy of the GRACE observations enabled us to adopt a new approach in the terms of the water balance equation. The new ET product derived in this study shows high similarity with three existing, high quality ET products across the CONUS, indicating the reliability of the approach. Since the new ET product is derived from observations, it can be regarded as a benchmark data set to evaluate the existing and new model-based ET products. Moreover, we downscaled the GRACE data with the aid of four LSMs to produce a continuous daily equivalent water thickness dataset with a spatial resolution of 0.125º and converted the USGS observed streamflow data to runoff depth. All the above products can serve as important hydro-meteorological data sets for assessment of hydrological and climatological changes, and evaluation of terrestrial water and energy cycle dynamics across the CONUS. These products will be

also valuable for studies and applications in drought assessment, water resources management, climate change evaluation, and so on.

Although this new ET product is derived from ground and satellite observations, there are several limitations with this approach and the product. Further study is needed to thoroughly address these limitations. First, the reconstructed ET from the water balance method is a basin-mean product and correspondingly has variable spatial resolutions depending on the area of each individual sub-basin. For example, the area of the 592 basins in this study ranges from 292 km^2 to 303,700 km^2. To produce gridded data, physical or statistical methods need be developed to disaggregate the areal-average ET to individual grid cells; the distributed hydrologic models and land surface models may be useful for this.

Second, the ET reconstruction method does not account for the impacts of water transfer in or out of the sub-basins by human activities such as irrigation and inter-basin water diversions; therefore, the ET estimates in these areas heavily impacted by these human activities may have higher uncertainty. We derived a map showing these sub-basins which have at least 10% of area controlled or affected by reservoirs and other human activities such as urbanization, mining, agricultural changes, and channelization using the USGS streamflow qualification codes for peak streamflow (http://nwis.waterdata.usgs.gov/nwis). 245 of the 592 basins have more than 10% area controlled by reservoirs, while 3 basins have more than 10% impervious cover due to urbanization, mining, agricultural changes, channelization, or other anthropogenic activities (Figure 3.9). These basins impacted by human activities are largely located in the Midwest (Figure 3.9). The stream flow interruption caused by human activities do not

affect the water balance-based ET reconstruction as long as no significant amount of water is diverted to another basin, because the ET in this study is derived on the basin level. However, the inter-basin transfer of water definitely can cause large errors in the water balance-based ET calculation. It is impractical for us to quantify the impact of the inter-basin transfer in this study due to lack of data. The general similar spatial patterns between ET derived in this study and the other three ET products from remote sensing and upscaled flux tower data in these basins impacted by human activities suggest that most of these basins do not experience substantial inter-basin transfer of water (Figure 3.4 and Figure 3.9).

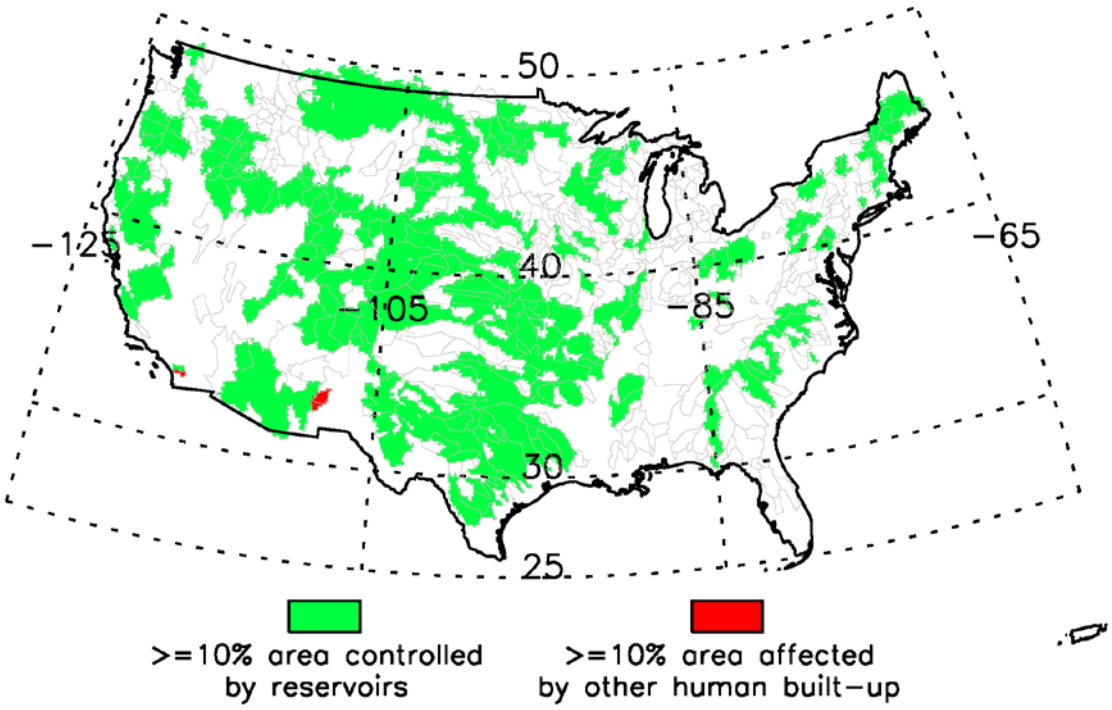


Figure 3.9 Locations of sub-basins are impacted by reservoirs and other human activity such as urbanization, mining, agricultural changes, and channelization.

Third, the ET estimate is directly calculated as the residual of all other water budget terms and inherits the measurement and processing errors in all other water budget terms. For example, some studies shows that the GRACE water thickness data can have an error of 2-3cm (Landerer and Swenson 2012). Although the evaluation of the PRISM precipitation shows a near zero bias over the CONUS but may have relatively larger errors in some regions (Daly et al. 2008). Finally, the availability of the ET reconstruction using this approach is limited by the availability of the measurements of the other water budget terms. However, the observation-based ET estimate in this study presents a best available ET estimate from the high quality observations. Therefore, there is strong reason to believe that this ET estimate is close to the "truth".

Chapter 4: Big data solutions enabled web GIS-based hydrological modeling framework for the conterminous United States

Hydrological modeling is widely applied in hydrologic research and practices through synthesis and simulation of hydrologic processes. Setting up and executing stand-alone hydrologic models can be difficult due to data limitation and lack of expertise. The objective of this study is to develop an automated and web accessible hydrological modeling framework for any non-hydrologists, as long as with internet access, who can organize hydrologic data, execute hydrologic models, and visualize results graphically and statistically for further analysis in real-time. By adopting Hadoop Distributed File System (HDFS) and Apache Hive, the efficiency of data processing and query was significantly increased. Two lumped hydrologic models, lumped Coupled Routing and Excess STorage (CREST) and Hydrological MODel (HyMOD), were integrated as a proof of concept in this web framework. Evaluation of selected basins over the conterminous United States (CONUS) were performed as a demonstration. Our vision is to simplify the processes of using hydrologic models for researchers and modelers, as well as to unlock the potential and educate the less experienced public on hydrologic models.

4.1 Introduction

As an important type of environmental modeling, hydrological modeling serves as an approach to conceptualize and investigate land surface and underground processes, which involves water and energy circulations, water resources management, and impacts of climate change (Leavesley 1994, Wood et al. 1997). Synthesis and simulation of any combinations of surface and groundwater processes are used as key tools in hydrological modeling to understand hydrologic processes by recurring historical events and predicting possible future occurrences (Viessman and Lewis 2003). Since forcing and output data from hydrologic models comprise heterogeneous geospatial data, desktop and web-based geographic information systems (GIS) are being used for data processing, analysis, visualization, and sharing (Bhatt, Kumar, and Duffy 2014, DeVantier and Feldman 1993). Some pilot studies have been conducted in this direction using desktop GIS with hydrological modeling. Van Der Knijff, Younis, and De Roo (2010) demonstrated the LISFLOOD model, which is a combination of a distributed hydrologic model with GIS for water balance study and flood simulations. Bhatt, Kumar, and Duffy (2014) introduced a modeling framework coupled Quantum GIS (QGIS) platform with Penn State Integrated Hydrologic Model (PSIHM) to provide effective functions including watershed delineation, simulation, and visualization.

However, in order to share modeling framework and collaborate with people distantly, it is inadequate to manage by stand-alone GIS based hydrological modeling framework; therefore, web technologies are adopted in such applications. Huang (2003) integrated TOPography based hydrological MODEL (TOPMODEL) into a web modeling and visualization system using web map service to allow user interacting with

environmental applications. The well-accepted Web-based Hydrograph Analysis Tool (WHAT) is another example using GIS to access real-time U.S. Geological Survey (USGS) streamflow data directly from web servers for visualization and validation (Lim et al. 2005). Comair et al. (2014) presented a GIS based hydrologic information framework to store and share hydrologic data, execute hydrologic model, and convey simulation results to stakeholders for better water resources management and decision-making. V. Boyina et al. (2015) implemented a hydrologic web-mapping application using GIS resources and meteorological observations to improve visualization and analysis process for hydrological modeling. Mantas, Liu, and Pereira (2015) created a web application for accessing earth observation data, which features online data visualization and analysis system, web services, and mobile applications.

Over the past decade, the advancement of satellite and ground observation technologies have been witnessed, which includes improvement on both spatial and temporal resolutions of these products. This improvement in return transforms spatiotemporal data sets in earth science studies into data sets with larger volume, higher update rate, and heterogeneous data types and formats. These kinds of revolution bring difficulties for web-based implementations of using these data due to time limit. One study disclosed that users of web applications are very sensitive to responding time in a matter of seconds (Galletta et al. 2004). With big data being involved in web-based hydrological modeling framework, solutions such as distributed file systems and high-performance computation are incorporated into such framework. Li, Yang, et al. (2013) developed a web-based system for high-performance analytics and visualization of big spatiotemporal data and climate model simulations. Hu, Cai, and DuPont (2015) coupled

an agent-based system with an environmental model and embedded it into a web application enabled by Hadoop-based high-performance computation to facilitate solving complex problems for watershed management.

Although previous studies pointed out different approaches of using GIS and web technologies with diverse types of modeling systems, none of them tried to create a general-purpose web-based hydrological modeling framework with big data solutions to cope with heterogeneous hydrologic data and models for non-hydrologist. In this study, a GIS enabled web-based framework is implemented to integrate different hydrologic models with the support of big data solutions. Hydrologic forcing data are initially organized, processed, and stored in the framework. By using HDFS and Apache Hive data infrastructure, the operation time for data processing, query, and input/output (IO) are reduced notably. Two lumped hydrologic models, lumped CREST and HyMOD rainfall-runoff models, are incorporated as a demonstration of such framework. Evaluations are conducted on selected basins over the CONUS. The following section demonstrates the implementation of this web framework and introduces the integrated hydrologic models. Sections 4.3 and 4.4 provide details of the data used in the framework and results from multi-basin evaluation and performance evaluation. Sections 4.5 and 4.6 are the discussion and conclusions related to this framework, respectively.

4.2 Method and Material

In this study, a web GIS-based hydrological modeling framework is created with HDFS and Hive to support data query, model execution, and results analysis and visualization, which consists of three components: data sources, servers with models, and user interface (Figure 4.1). The next sub-section introduces the structure of this web framework and two integrated hydrologic models will be described in Section 4.2.2.

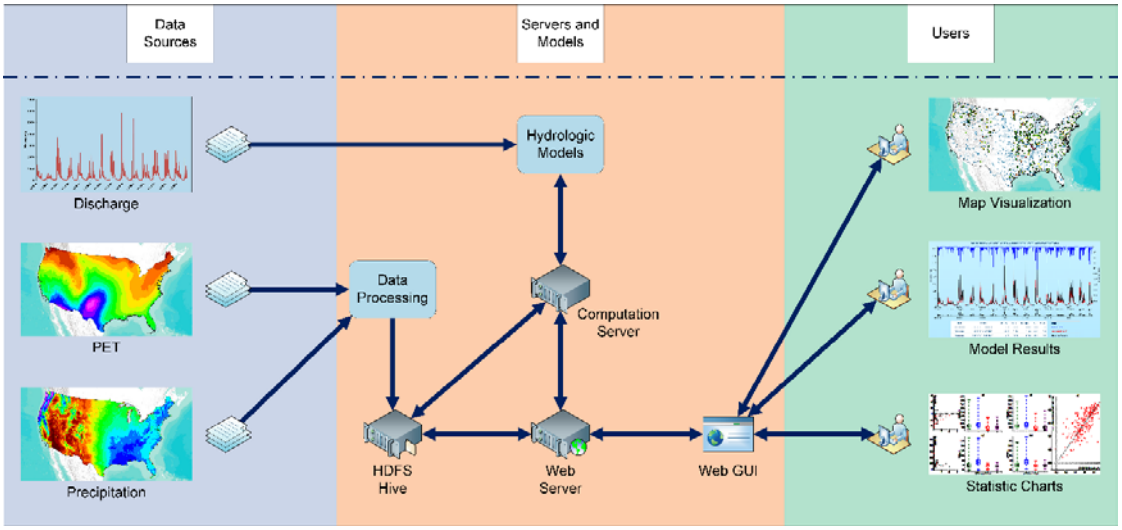


Figure 4.1 The architecture of the web GIS-based hydrological modeling framework.

4.2.1 Web Framework Implementation

Data sources

USGS discharge data are used for hydrologic models calibration and validation. The data are downloaded programmatically using USGS water services (http://waterservices.usgs.gov/) as tab-separated values, which are ready to use in web environment. To execute hydrological models, such as lumped CREST and HyMOD, precipitation data and potential evapotranspiration (PET) data are required as forcing data

in the same spatial and temporal resolutions. These gridded data are usually obtained from satellite observations or gauge based interpolation in the form of a data array containing data cells in both rows and columns mapping to corresponding longitude and latitude for each time step. When using file based method for data retrieving and calculation, most of the running time is consumed during the file opening and loading to memory phases if thousands of files needs to be opened and processed at the same time (e.g. loading and processing more than 3650 data files for fine-resolution CONUS wide 10-year daily precipitation data). To the best of our knowledge, data query and management will also be problematic with gridded data files because no effective query language or other techniques have been published to work directly with large amount of data files. A possible workaround is to convert the data files, transfer them to HDFS, and store them into Hive.

Server structure

Apache Hadoop, which is an open source programming framework, consists of MapReduce and HDFS to provide solutions to store, process, and manage large data sets (Borthakur et al. 2011, Polato et al. 2014). MapReduce is an architecture to divide data and tasks (“map” module) and run them with parallelization in a distributed computing environment (“reduce” module) (Bhosale and Gadekar 2014, Hu, Cai, and DuPont 2015). HDFS is a file storage system built on top of Hadoop framework to facilitate high performance file services with high reliability and scalability (Borthakur et al. 2011, Shvachko et al. 2010). As another component of Hadoop framework, Apache Hive serves as a data warehouse infrastructure that structures data into common database paradigm

and provides a language similar to Structured Query Language (SQL) for data query and management (Thusoo et al. 2010). In this study, precipitation data are in binary format and PET data are in Hierarchical Data Format (HDF) format originally. However, both of them have the same daily temporal resolution and a similar structure that contains metadata and array-like data values. The precipitation and PET data originally have different spatial resolutions and they are projected to 0.125º before they are used in the web framework. Converting the gridded precipitation and PET data to comma-separated values (CSV) format is straightforward. Each row of data in CSV file represents one year of data for a specific location (i.e. a data cell in the gridded file). Although, the size of CSV file is larger than the same amount of data stored in binary or HDF format, CSV is one of the standard format to transfer data to Hive data infrastructure and the size of data will be compressed after stored in Hive. The coordinate index and the year number are inserted to each data row as a unique identifier of that row. The coordinate index is calculated as in the following equation:

$$Coordinate\ Index = \frac{(90 - cLat) \times 360}{dRes^2} + \frac{180 + cLon}{dRes} \tag{4.1}$$

where cLat and cLon are the latitude and longitude of a given grid cell, respectively and dRes is the data resolution. All the given values and variables in the equation above have a unit of degrees; dRes is set to 0.125 in the study. As a result, it is possible to convert coordinates of every grid cell of any data with any resolution in this way and it is reversible to determine the coordinates of a grid cell from the coordinate index. This mechanism of data format ensures that the web framework is scalable to accommodate any gridded data with any spatial and temporal resolution by design. Converted

precipitation and PET data are transferred to HDFS and imported into Hive by web Application Programming Interfaces (API).

Implementation and user interface

A five-node computer cluster is configured as a testbed for the web framework, including three nodes for HDFS and Hive, one node for data calculation and model execution (computation server), and one node for website deployment (web server). Apache HTTP Server is deployed on the web server to host the web portal of this hydrological modeling framework and handle all communications between users and other servers. Hydrologic models are converted to Python programming language (Rossum 1997) and capsulated as an individual web program with standardized data input and output. The converted hydrologic models are invoked through FastCGI (Fast Common Gateway Interface), which is a protocol defined for communications between servers and FastCGI programs (Adida 1997). The web portal is built with basic hypertext markup language (HTML) and JavaScript and the results are visualized with open-source JavaScript libraries.

Users are provided with several ways to select a basin of interest from the web portal (Figure 4.2). Once a basin is selected, the web server is going to communicate with the computation server to search for calculated areal mean precipitation and PET data for the selected basin. If the basin has been previously calculated by another or the same user, the current user will be notified to proceed to the next step. If the basin is never selected for simulation, precipitation and PET data on Hive will be queried and retrieved for the

selected basin. The retrieved data are then sent to computation server and areal mean precipitation and PET are calculated for each time step:

$$\overline{Data} = \frac{\sum(Data_i \times a_i)}{\sum a_i} \quad (4.2)$$

where $\overline{Data}$ is the areal mean precipitation or PET; $Data_i$ refers to precipitation or PET data value of the ith grid cell in a given basin and a_i is the area of the ith grid cell. Once

the areal mean values are calculated, it is stored on the computation server for repeated use.

(a)

USGS Gauge List

Show 10 entries Search:

Site No.	Site Name	HUC	Drainage Area (sq mi)
01021000	St. Croix River at Baring, Maine	01050001	1374
01034500	Penobscot River at West Enfield, Maine	01020005	6422
01046500	Kennebec River at Bingham, Maine	01030003	2715
01054000	Androscoggin River near Gorham, NH	01040001	1361
01059000	Androscoggin River near Auburn, Maine	01040002	3263
01092000	MERRIMACK R NR GOFFS FALLS, BELOW MANCHESTER, NH	01070002	3092
01100000	MERRIMACK RIVER BL CONCORD RIVER AT LOWELL, MA	01070002	4635
01129500	CONNECTICUT RIVER AT NORTH STRATFORD, NH	01080101	799
01144500	CONNECTICUT RIVER AT WEST LEBANON, NH	01080104	4092
01170500	CONNECTICUT RIVER AT MONTAGUE CITY, MA	01080201	7860
Site No.	Site Name	HUC	Drainage Area (sq mi)

Showing 1 to 10 of 323 entries Previous 1 2 3 4 5 ... 33 Next

(b)

Close

Figure 4.2 Web interface of the proposed modeling framework and its options for users to select their basins of interest in the framework: (a) selection from the map by clicking the basin's corresponding gauge point, (b) selection from a list of gauges or by searching gauge information (as shown in the red rectangle in Figure 4.2(b)).

Multiple options are provided for date range selection, model selection, and model parameters input (Figure 4.3). Data for executing the models are currently provided between January 2000 and December 2013. Default parameter values for both lumped

CREST and HyMOD models are provided for selected basins. These values are retrieved by model calibration using Markov Chain Monte Carlo (MCMC) parameter optimization method (Metropolis et al. 1953) for the time period from 2001 to 2005 inclusive; right after a warm-up time period of the year of 2000 to abate the uncertainty in initial conditions and balance the soil state in models (Xue et al. 2015). The maximum Pearson correlation coefficient (CC) is used as single objective function to perform automatic hydrologic calibration. These model parameters are adjustable within given ranges and detailed information is provided in the next sub section.

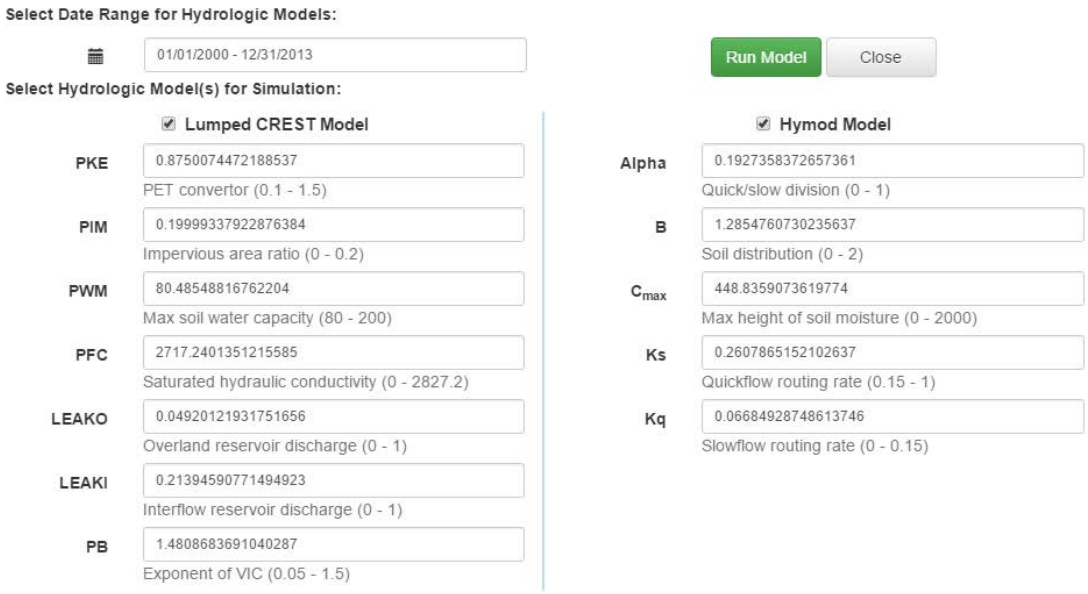


Figure 4.3 User input box for date range and model parameters.

Model execution and results visualization are automatic after user issues command to the web server. The results are visualized for each selected hydrologic model (Figure 4.4), including a hydrograph displaying plots of observed and simulated discharge and precipitation rates versus date, a panel showing statistical metrics for user defined, calibration, and validation time periods separately, and functionalities, such as zooming-

in to a time period and mouse-over hydrograph for instantaneous values of observed and simulated discharge, precipitation, and date.

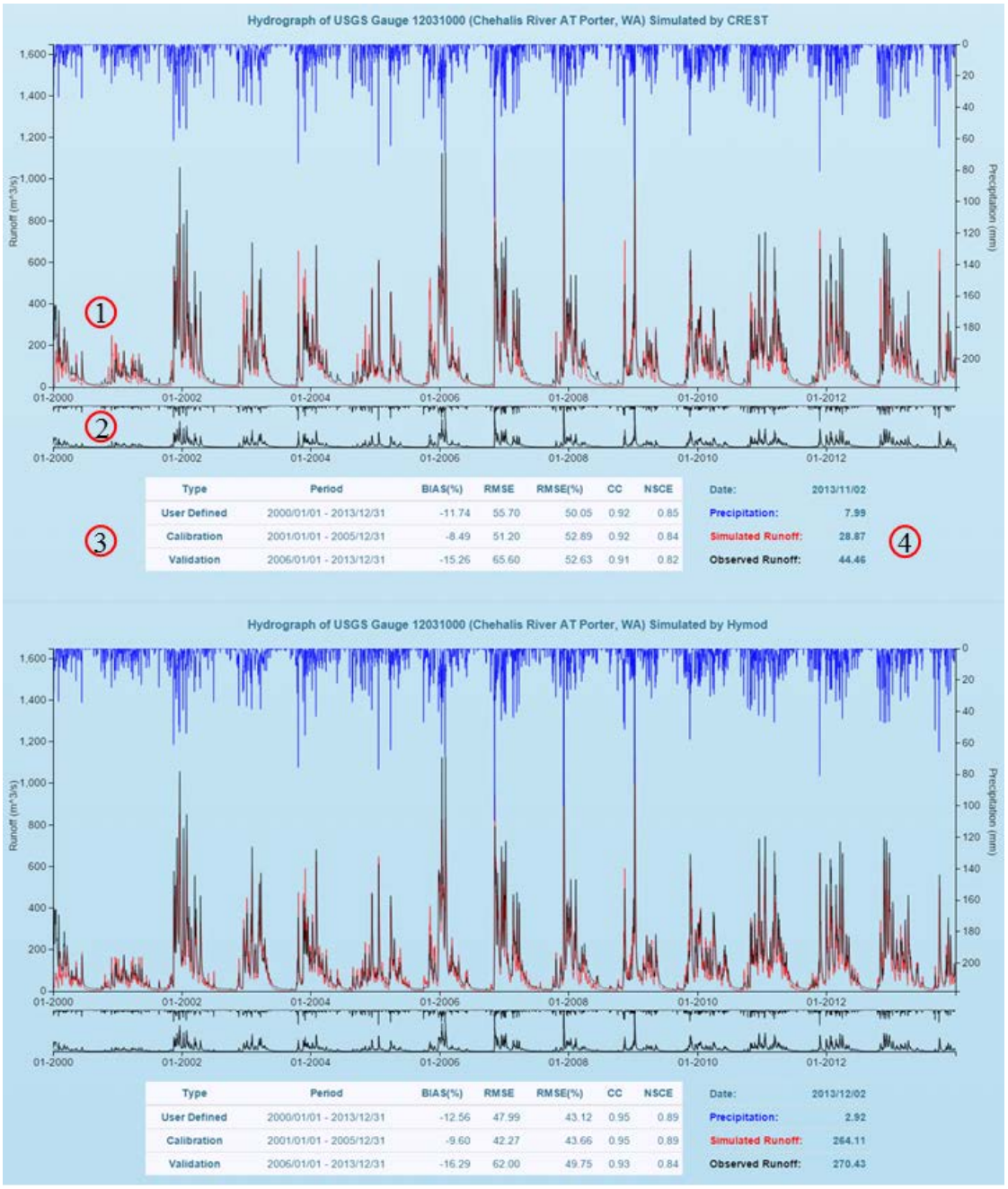


Figure 4.4 The results from executing both lumped CREST model (top panel) and HyMOD model (bottom panel) for a selected basin. Each panel contains four sections: (1) hydrograph section, (2) zoom-in section, (3) statistics section, and (4) mouse-over value section.

4.2.2 Hydrologic Models

This web GIS-based hydrological modeling framework is designed to accommodate any hydrologic models. Two commonly used lumped hydrologic models, lumped CREST and HyMOD models, are integrated into the framework as a proof of concept. Lumped hydrologic models treat the entire basin as a whole piece and no water is routed inside the target basin. Both of these two hydrologic models employ the same precipitation and PET data as their forcing data.

Lumped CREST Model

The lumped CREST (Coupled Routing and Excess Storage) model is a simplified version of grid based distributed CREST model (Wang et al. 2011, Xue et al. 2013), which was developed by the University of Oklahoma (http://hydro.ou.edu) and NASA SERVIR Project Team (http://www.servir.net). Lumped CREST computes the runoff generation components (e.g., surface runoff and infiltration) using the variable infiltration capacity (VIC) which is a concept originally presented in the Xinanjiang Model (Zhao et al. 1980, Zhao 1992) and later implemented in the VIC Model (Liang et al. 1994). After precipitation reaches the ground, some of the water evaporates into the atmosphere, whereas infiltration is split out of the excess rainfall through VIC curve. The rest of the excess rainfall is separated into overland excess rainfall (R_O) and interface excess rainfall (R_I) by using the saturated soil hydraulic conductivity (Wang et al. 2011, Zhang et al. 2014). R_O and R_I are further routed through overland and interflow linear reservoirs respectively, which are controlled by the corresponding discharge multipliers. The summation of overland flow and interflow forms the total simulated discharge. Seven

parameters in lumped CREST model are provided to accept user input and used for model calibration. The descriptions and empirical ranges are given in Table 4.1 (Xue et al. 2013).

Table 4.1 Parameters of lumped CREST model.

Parameter	Description	Range
PKE	Multiplier to convert PET to actual evapotranspiration (AET)	0.1 – 1.5
PIM	Impervious area ratio	0– 0.2
PWM	Maximum water capacity of the soil layer (mm)	80 – 200
PFC	Saturated soil hydraulic conductivity (mm / day) to separate excess rain	0 – 2827.2
LEAKO	Overland reservoir discharge multiplier	0 – 1
LEAKI	Interflow reservoir discharge multiplier	0 – 1
PB	Exponent of VIC curve	0.05 – 1.5

HyMOD model

HyMOD model is a conceptual lumped hydrologic model for basin wide hydrologic simulation based on the Probability Distributed Model (PDM) (Wagener et al. 2001, Moore 2007). This model consists of several quick release reservoirs and one parallel slow release reservoir operating at the same time (Zhang et al. 2013). Precipitation is first used to fill the water storage capacity and yield the AET, and then the excess rainfall is separated into quick and slow flows by the parameter Alpha. These quick and slow flows are routed through corresponding quick and slow release reservoirs with rates governed by rate parameters of quick and slow release reservoirs, respectively. The summation of quick and slow flows is the total simulated discharge. Five parameters in HyMOD model are provided to accept user input and used for model calibration. The descriptions and empirical ranges are given in Table 4.2 (Quan et al. 2015, Zhang et al. 2013, Herman, Reed, and Wagener 2013).

Table 4.2 Parameters of HyMOD model.

Parameter	Description	Range
C_{max}	Maximum storage capability in basin (mm)	0 - 2000
B	Soil distribution parameter	0 – 2
Alpha	Quick/slow routing division parameter	0 – 1
Kq	Quick release reservoir rate parameter (day)	0.15 – 1
Ks	Slow release reservoir rate parameter (day)	0 – 0.15

4.3 Data and Study Area

4.3.1 Data

Discharge data, precipitation data, and PET data are imported into this web framework for a test run. As previously mentioned, discharge data are downloaded programmatically from USGS and they are daily mean discharge observations from gauge stations between January 2000 and December 2013. The units of USGS discharge data are originally cubic feet per second (cfs) and converted to cubic meter per second (cms).

The precipitation data are produced by the PRISM (Parameter-elevation Regressions on Independent Slopes Model) group at Oregon State University (http://www.prism.oregonstate.edu). The product is daily, 4 km gridded estimate of precipitation for the CONUS based on observations from a wide range of surface stations with quality control and corrections (Daly et al. 2008). The PRISM interpolation method computes climate elevation regression for each grid cell, and stations entering the regression are assigned weights based primarily on the physiographic similarity of the station to the grid cell. The PRISM data is the official spatial climatological data of U.S. Department of Agriculture (USDA). In this web framework, all analyses were conducted with a spatial resolution of 0.125^{o}. Therefore, the PRISM precipitation data were first aggregated from 4 km to 0.125^{o} before imported and used in the web framework.

The PET data were extracted from a global terrestrial evapotranspiration (ET) product, which adopted satellite remote sensing-based algorithms using a modified Penman-Monteith approach with normalized difference vegetation index (NDVI) based biome-specific canopy conductance to estimate canopy transpiration and soil evaporation

and using a Priestley-Taylor approach to estimate open water evaporation (Zhang et al. 2015, Zhang et al. 2009, Zhang et al. 2010). The PET data are daily products, which were first aggregated from 8 km to 0.125º and then imported and used in the web framework.

4.3.2 Study Area

USGS gauge stations over the CONUS were used to perform an evaluation using this web framework. Wan et al. (2015) utilized a screening process to select gauge stations and a similar approach was adopted in this study. First, all selected gauge stations should have continuous daily discharge data between January 2000 and December 2013 and the data should be actual observation values instead of estimate values as well as not being disturbed by ice or other natural factors. Second, differences between USGS provided drainage areas of gauge stations and the areas derived from the 0.125º geographic grid should be less than 20%. Third, the drainage area of each gauge station should be larger than two 0.125º grid cells. After the screening process, only 323 gauge stations were chosen for further analysis (Figure 4.5).

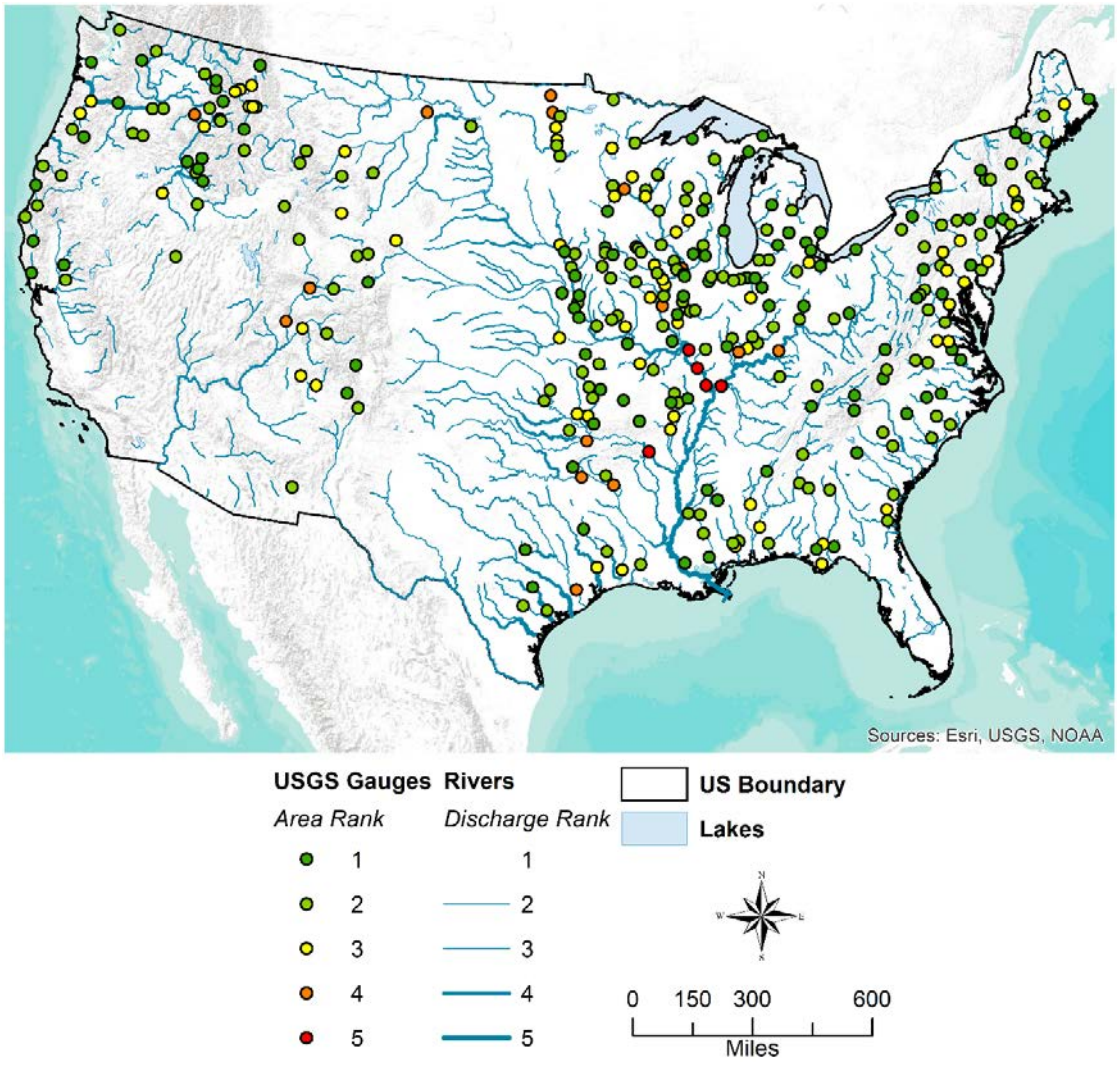


Figure 4.5 Distribution of 323 selected USGS gauge stations and major river channels. The gauge stations are color coded by gauge areas using geometrical interval classification and river channels are classified by discharge rate (level 1 of river channels was intentionally assigned blank legend to reduce displayed river channels in the figure). From level 1 to 5, gauge controlled area and river discharge rate gradually increase.

4.4 Evaluations and Results

4.4.1 Multi-basin Evaluation

Multi-basin evaluation was designed to evaluate the functionalities of this web framework by sequentially executing web accessible lumped CREST and HyMOD models for each of the selected basins in a programmatic way with the same forcing data and time ranges. Nash-Sutcliffe coefficient of efficiency (NSCE), CC, normalized bias (BIAS (%)), and normalized root mean-squared error (RMSE (%)) were used as the statistical metrics to quantify model performance. NSCE values range from negative infinity to one, where one indicates perfect agreement between observed and simulated discharge data. The agreement decreases as NSCE values distance from one. CC values range between -1 to 1 inclusive and is used to measure the degree of linear relationship between observed and simulated discharge data with 1 being total positive correlation, 0 being no correlation, and -1 being total negative correlation. The normalized bias assesses the relative difference between observed and simulated discharge data and normalized RMSE measures the mean error magnitude between these two data sets. Both normalized bias and RMSE favor of the optimal value of 0%. The equations for calculations of NSCE, CC, BIAS (%), and RMSE (%) are given in Equations (4.3) to (4.6), respectively.

$$\text{NSCE} = 1 - \frac{\sum(Obs_i - Sim_i)^2}{\sum(Obs_i - \overline{Obs})^2} \tag{4.3}$$

$$\text{CC} = \frac{\sum(Obs_i - \overline{Obs})(Sim_i - \overline{Sim})}{\sqrt{\sum(Obs_i - \overline{Obs})^2 \sum(Sim_i - \overline{Sim})^2}} \tag{4.4}$$

$$\text{BIAS (\%)} = \frac{\sum Sim_i - \sum Obs_i}{\sum Obs_i} \times 100 \tag{4.5}$$

$$\text{RMSE}\ (\%) = \frac{\sqrt{\frac{\sum(Obs_i - Sim_i)^2}{n}}}{\overline{Obs}} \times 100 \tag{4.6}$$

where Obs and Sim denote observed and simulated discharge data respectively; $\overline{Obs}$ and $\overline{Sim}$ means the arithmetic mean of observed and simulated discharge data separately; i is the ith values of observed or simulated discharge data; n is the total number of pairs of observed and simulated discharge data.

As aforementioned, automatic calibrations are applied for each basin using MCMC optimization method for the time period between 2001 and 2005 inclusive. The time period between 2006 and 2013 inclusive is used as validation period in this evaluation.

Evaluation results are illustrated in this section and all the results are calculated and visualized by using this web framework. Figure 4.6 shows the distributions of statistical metrics of lumped CREST and HyMOD models with and without calibration for calibration and validation time periods. In general, both hydrologic models perform similarly to each other for both calibrated and uncalibrated cases during both time periods. The bottom, middle, and top lines of the box represent the 25th percentile, median, and 75th percentile respectively, while the bottommost and topmost lines show the minimum and maximum values respectively. It is apparent that the distribution of NSCE shifts to higher values from uncalibrated results to those calibrated results (Figure 4.6a, b). Moreover, more than 75% of NSCE values are positive after calibration for both hydrologic models in both time periods. Likewise, the distribution of CC in the calibrated results also shifts to higher values with almost all positive values (Figure 4.6c and d). In addition, the mean and spread of normalized RMSE and bias decrease considerably after

model calibration is applied (Figure 4.6e, Figure 4.6f, Figure 4.6g, and Figure 4.6h). The similarity and changes in all statistical metrics indicate that both hydrologic models perform comparably and the automatic calibration method that was used to provide optimal parameters for both hydrologic models substantially improves model results in the web framework.

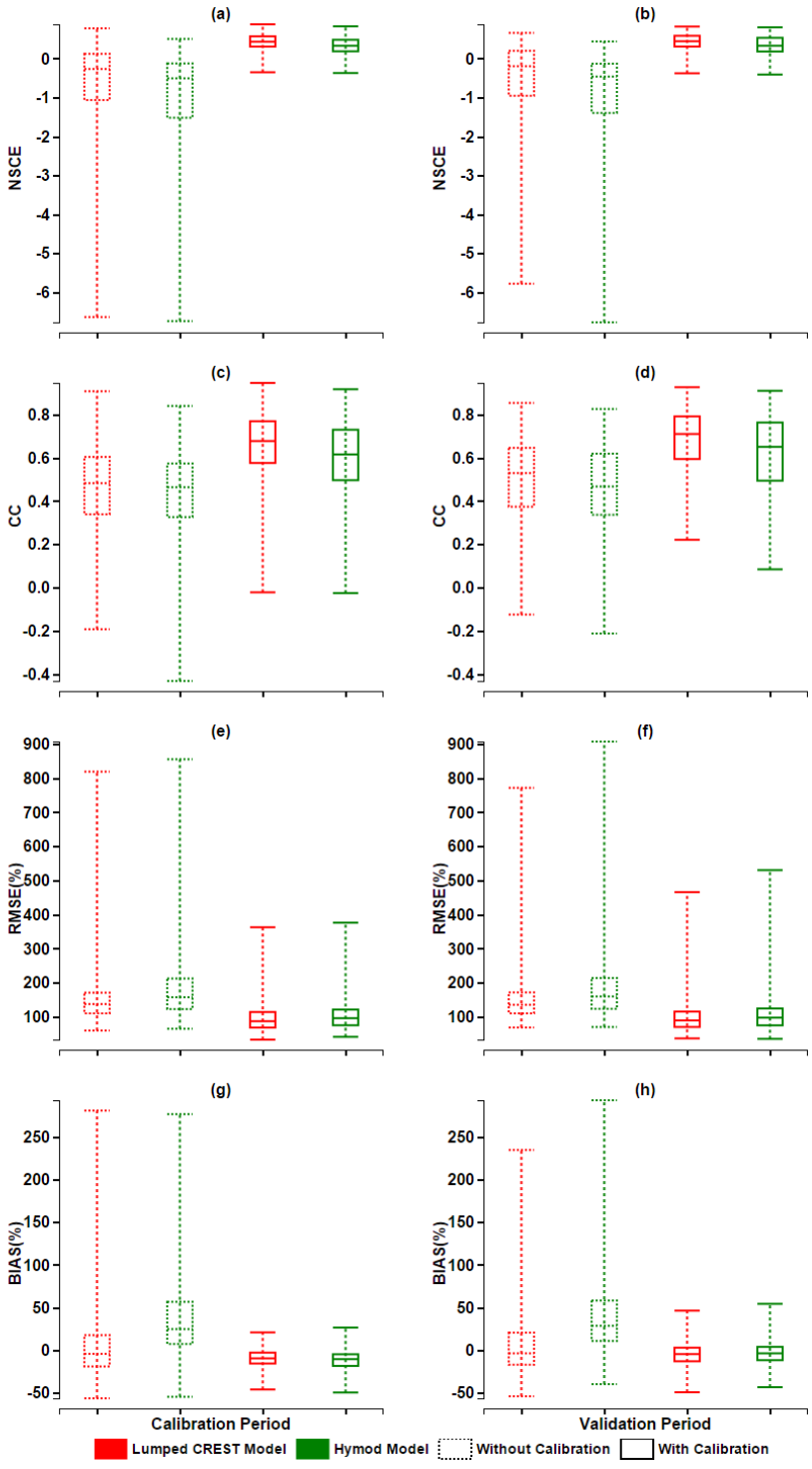


Figure 4.6. Comparison of statistical indices (NSCE, CC, RMSE (%), BIAS (%)) between Lumped CREST and HyMOD models before and after calibration during calibration and validation time periods.

Further evaluations are conducted using model results after calibration to reveal more details. Figure 4.7 shows inter-comparisons between calibration and validation time periods for both lumped CREST and HyMOD models using CC and NSCE values calculated from observed and simulated discharge. The dotted vertical and horizontal lines divided each plot into four quadrants to separate positive and negative values. The dotted diagonal is the bisector of the first quadrant. Both models present high correlations between calibration and validation time periods for both CC and NSCE values. It is clear that only a small number of basins have negative values in CC and NSCE. HyMOD model presents a slightly higher correlation than lumped CREST model in both scenarios.

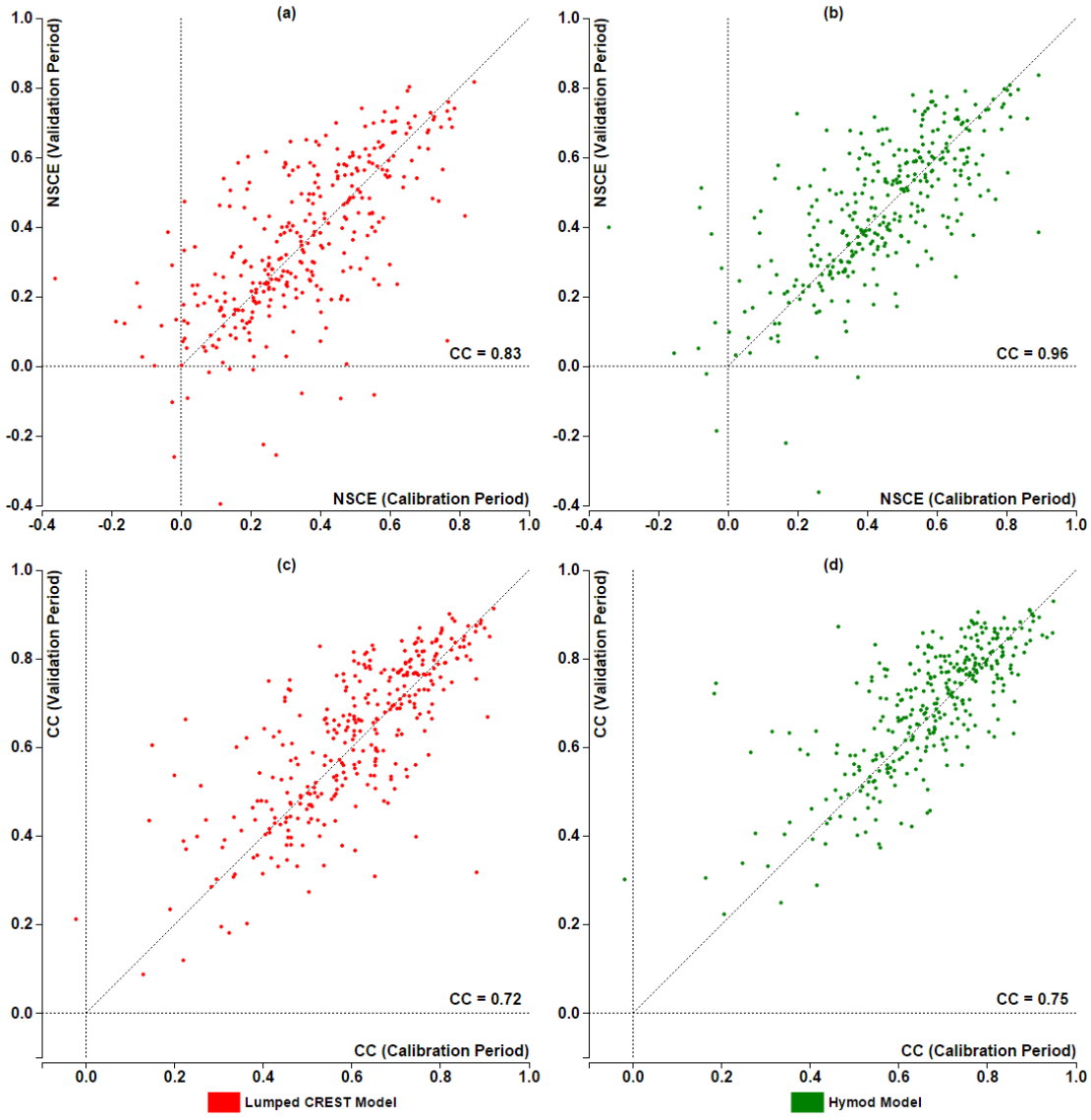


Figure 4.7 Comparison of statistical metrics (CC and NSCE) between Lumped CREST and HyMOD models after calibration between calibration and validation periods.

4.4.2 Performance Evaluation

To facilitate fast data query, Apache Hive was used with the support of HDFS. A performance evaluation was developed to compare the execution time of SQL queries on both PostgreSQL and Hive. PostgreSQL is a commonly used relational database

management system (Herzog 1998). Queries were executed on both PostgreSQL and Hive against data with the same structure and volume, which comprised about 1.3 million converted PET data records over the CONUS. Pseudocode of the queries are listed in Table 4.3. Query 1 was a full table scan for the total number of data records. Queries 2 and 3 were used to retrieve partial and all the data from table by using a year number as the filter and no filter, respectively. Queries 4 to 8 were filtering data retrieval process by using basin indices. “Basin_Guage_ID” used in Table 4.4 is an array of coordinate indices calculated by Equation (4.1) for target basins. For instance, if a basin covers an area of five grid cells which have corresponding coordinate indices of 1149854, 1152734, 1155614, 1155615, and 1155616, the query is written as “Select * from table where index in (1149854, 1152734, 1155614, 1155615, 1155616)”. Five basins used in Queries 4 to 8 ranged from small to large with an area of 1516, 5926, 28635, 158138, and 713200 square miles (sq mi), and these basins were selected based on geometrical interval classification of the area of all 323 selected basins.

Table 4.3 Pseudocode and description of database queries.

	Pseudocode	Description
Query 1	Select count (*) from table;	Full table scan
Query 2	Select * from table where year = 2005;	Select partial data
Query 3	Select * from table;	Select all data
Query 4	Select * from table where index in Basin_03345500;	Select data from a basin with an area of 1516 sq mi
Query 5	Select * from table where index in Basin_07185000;	Select data from a basin with an area of 5926 sq mi
Query 6	Select * from table where index in Basin_03377500;	Select data from a basin with an area of 28635 sq mi
Query 7	Select * from table where index in Basin_07263450;	Select data from a basin with an area of 158138 sq mi
Query 8	Select * from table where index in Basin_07022000;	Select data from a basin with an area of 713200 sq mi

The results of the performance evaluation are illustrated in Figure 4.8. Each query was executed consecutively on PostgreSQL and Hive for ten times separately. Maximum and minimum execution times were removed before calculating the mean execution time for each query. The reason for conducting the removal of max/min values was that the database system sometimes needs warm-up time for hard drive-memory data exchange, especially when the database is queried for the first time. Figure 4.8 shows a clear increase of query execution time in PostgreSQL as a function of the number of retrieved data records for the same type of queries, whereas the query execution time of Hive stays almost consistent for all queries in this evaluation. All the queries executed on Hive are

faster than these executed on PostgreSQL. Queries 6 to 8 executed on Hive were more than ten times as fast as these executed on PostgreSQL.

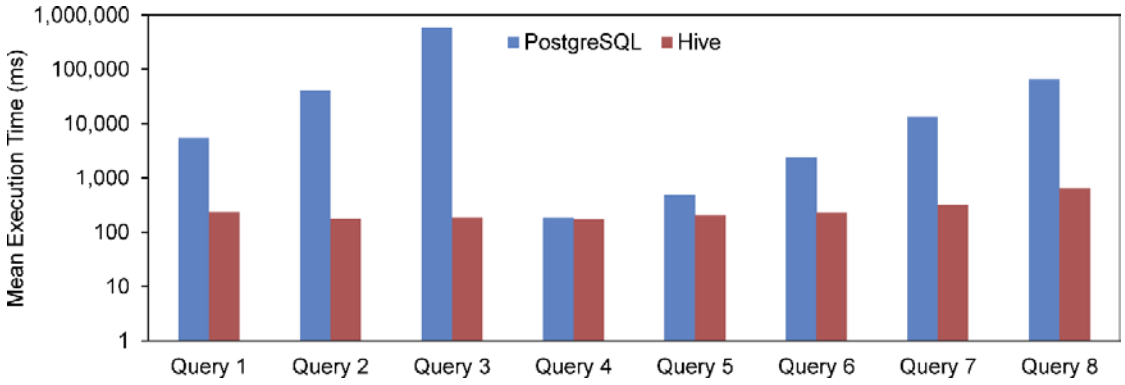


Figure 4.8 Comparison of mean execution time between PostgreSQL and Hive for eight different SQL queries.

4.5 Discussion

Although this web GIS-based hydrological modeling framework does not directly improve hydrologic models, it is designed to incorporate any hydrologic models or other type of models, such as geotechnical model SLIDE (Hong et al. 2015), through a web accessible environment to better serve hydrologic research and education. From research perspective, researchers can use this framework to select their basin of interest, time period, models, and parameters to execute models and attain results for evaluation and validation. From education perspective, this web framework provides pervasive usability for anyone with Internet access, making it possible for people without hydrology background to learn practical hydrological modeling and evaluation skills. With this plug-in free web framework, users only need a modern web browser to dive in instead of taking time installing software, downloading data, or configuring models.

4.5.1 Big data support

Big data refers to any collection of data sets with large size, high complexity, and fast growth rate, which are usually difficult to be managed and processed by traditional databases or file systems (Kitchin 2013, Bhosale and Gadekar 2014, Fernández et al. 2014, Vitolo et al. 2015). Initially, the precipitation and PET data used in this study are file based. The precipitation data are stored as binary files and each file contains daily data for one day, whereas the PET data are wrapped as HDF files and each file contains daily data for one year. It is also possible that other data formats will be used in the future when more models are integrated with different forcing data. With inconsistent file formats and structures, it is difficult to retrieve, query, and process these data in an online

environment, especially when thousands of data files need to be load and read for long time series.

Big data solution is integrated to solve these problems by using HDFS and Hive. HDFS provides reliable and elastic distributed file storage and management, and Hive is capable of accommodating large data sets for high performance data query and processing. Another advantage of using Hive is that data will be highly compressed and data query speed will be significantly increased if a table is created and stored as ORC (Optimized Record Columnar) format (Huai et al. 2014). Data in this study were stored as ORC format which yielded approximately 85% decrease in data size and five times speedup in data query when comparing to the same data stored in regular Hive table. These characteristics of big data solution guarantee the ability of expanding this web framework to host much more data in the future.

4.5.2 Data and Models Trade-off

The precipitation and PET data originally have a spatial resolution of 4 km and 8 km, respectively and both of them were aggregated to a lower spatial resolution of 0.125^0 before imported into this web framework. The major reason for downscaling forcing data to lower resolution is that this framework is not only established for applications in the CONUS, but also for global applications, especially in the data sparse areas/countries. For example, precipitation or PET data are only available at coarse spatial resolutions for some parts of Africa. In order to use this web GIS-based modeling framework globally, it is wise to adopt data with relatively fine spatial resolution as a test run and gradually migrate to high spatial resolution for data rich areas.

For this framework, it is possible to incorporate both lumped and distributed hydrologic models. Distributed hydrologic models are capable of maintaining spatial variabilities of basins, whereas model complexity needs to be considered to balance computation time and resolution of forcing and output data (Carpenter and Georgakakos 2006). Theoretically, with better understanding of the dynamics of hydrologic processes, rapid development of computation technologies, and growing availability of fine-resolution forcing data, the performance of distributed hydrologic models should exceed lumped hydrologic models. However, multiple studies concluded that distributed hydrologic models did not always outperform lumped hydrologic models due to uncertainties in hydrologic models, observation data, spatial characteristics of basins, and so forth (Grayson, Moore, and McMahon 1992, Liu and Gupta 2007, Reed et al. 2004, Khakbaz et al. 2012). For the simplicity of illustration, lumped CREST and HyMOD models were selected in this study rather than distributed hydrologic models. In addition, with the increase of spatial resolution of the forcing data, the computational efficiency of distributed models will decline much more than that of lumped models. As a result, lumped model will benefit the future studies when the web framework performance is examined using data with high spatial resolution.

4.5.3 Scalability

This web framework is created with scalability in mind in several aspects. First, this framework used free or open source software for implementation, including data infrastructure (Apache Hive), file system (HDFS), web server (Apache HTTP server), Python libraries (Numpy for calculation, h5py for HDF-related operation, Spotpy for

parameter optimization, and et cetera), and JavaScript libraries and extensions (Bootstrap for framework structure, Leaflet for map integration, D3 for general purpose visualization, and et cetera). For detailed information regarding aforesaid software, refer to Steiniger and Hunter (2013), Zavala-Romero et al. (2014), and Swain et al. (2015). With free or open source software, maintenance and further development of this framework is possible. Second, models integrated in this framework are not limited to the demonstrated ones. If a model supports gridded forcing data and can be directly used or recoded to be compatible with FastCGI, the model can be integrated into this web environment. Execution time of a new model should be evaluated beforehand because web applications are highly time sensitive. Third, data storage is optimized to incorporate data with different spatial and temporal scales by using coordinate index and adjusting data table structure. Fourth, since the entire web GIS-based hydrological modeling framework is built on top of big data enabled distributed file system and data infrastructure, the framework can be transferred to a computer cluster with any number of nodes.

4.6 Conclusion

The web GIS-based hydrological modeling framework proposed in this study, with big data support and modeling integration, provides a general purpose web accessible infrastructure for data storage, processing, hydrologic models execution, as well as graphical and statistical results visualization and evaluation. By adopting HDFS and Hive, the time consumption for processing and querying data declines drastically. Two lumped hydrologic models, lumped CREST and HyMOD rainfall-runoff models, were integrated in this web framework as a demonstration. 323 basins were selected over the CONUS to conduct the multi-basin evaluation. Both hydrologic models presented significant performance increase after automatic calibration. The statistical results for both models after calibration showed noticeable similarity between calibration and validation time periods. The objective is to facilitate the processes of using hydrologic models for researchers as well as the public and bridge the gap between complicated hydrological modeling education and the studious non-hydrologist.

Future expansion of this web GIS-based modeling framework will incline towards three directions. For CONUS wide web modeling simulation, the framework will be deployed on more powerful cluster to accommodate data with higher spatiotemporal resolutions. Temperature data will be added to the framework for enabling snow module for both lumped CREST and HyMOD models. For web modeling simulation in data sparse areas/countries, such as regions of Africa, this framework will be examined for usability. For engaging different types of models, geotechnical SLIDE model will be added as an experiment. Given the elasticity of this web framework, it is conceivable that

different types of data and models could be incorporated for local, regional, and global wide simulations in various fields of studies.

Chapter 5: Overall Conclusion and Future Work

5.1 Summary

Hydrological modeling has been widely used in hydrologic events predictions, water resources management, climate change evaluation, and hydrology education. The performance efficiency and results accuracy of hydrological modeling are always crucial to the users. Furthermore, sharing hydrological modeling framework, including the data, the models, and the results is the key to collaboration with researchers and non-hydrologists to conduct research, solve problems, and educate people with interest in hydrology. As it is challenging to achieve the overarching goal by stand-alone hydrological modeling systems, a novel cyberinfrastructure is established in this dissertation to coordinate hydrologic data collection and organization, support hydrologic model integration and execution, improve model results aided by analytics and visualization, and share the whole framework to anyone with Internet access.

An approach of combining cloud-computing service with crowdsourcing method is employed to establish a cyberinfrastructure for flood events collection, on-demand, location-based visualization, and statistical analysis. It creates a network that could involve citizen-scientists participation, allowing the public to submit personal accounts of flood events to help the flood disaster community to archive detailed information of flood events, investigate past flood events, and get prepared for forthcoming flood events. This cyberinfrastructure delivers an opportunity to modernize the existing methods utilized by the flood disaster community ultimately to collect, manage, visualize, and analyze data with flood events.

A new actual ET product across the CONUS is derived from high quality ground and satellite observations, including the PRISM precipitation data, USGS observed discharge data, and GRACE EWT data that has been downscaled using land surface models with an innovative and unique bias correction algorithm. This data set covers 73% of the CONUS and is available for eleven years. The CONUS-wide, decadal, and continuous monthly ET estimates are calculated by using water balance equation enabled by the wide availability and accuracy of the GRACE observations. The new ET product derived in this research shows high similarity with three existing, high quality ET products, indicating the reliability of the approach, which can be regarded as a benchmark data set to evaluate other ET products over the CONUS. Moreover, the downscaled the GRACE data, converted USGS observed runoff depth data, and the new reconstructed ET product can serve as important and valuable data sets for assessment in hydro-meteorological studies and applications.

Lastly, a web GIS-based hydrological modeling framework is proposed with big data support and modeling integration to provide a general-purpose Internet accessible modeling infrastructure for hydrologic data processing and management, hydrologic models execution, and graphical and statistical results visualization and evaluation. By adopting big data solutions (i.e. HDFS and Hive), the efficiency for processing and querying data is increased considerably. Two lumped hydrologic models, lumped CREST and HyMOD rainfall-runoff models, are integrated in this web framework as a proof of concept. 323 basins has been selected with strict screening over the CONUS to perform the multi-basin evaluation. Both hydrologic models present significant improvement after automatic calibration. The statistical results for both models after calibration show

noticeable similarity between calibration and validation time periods. This web-based sharable hydrological modeling framework is developed for both researchers and non-hydrologists and is believed to facilitate the processes of preparing and archiving hydrologic data, executing hydrologic models, and conducting analysis and visualization for hydrologic research and education.

5.2 Limitations and Future Work

The cloud service used in the flood disaster cyberinfrastructure is a free service provided by Google Fusion Table (GFT). This free cloud service has limits, such as maximum size of each Fusion Table, maximum capacity of each Google cloud account, maximum insertion size, maximum requests per day, and maximum response size per query. Some constrains have been resolved when GFT upgrades from version 1.0 to 2.0 (Google 2015b). However, if other limits have been reached, it is always possible to find other paid cloud services. The flood cyberinfrastructure will be linked to real-time and archived ground and satellite observations, as well as model-simulated results, which will be beneficial as a validation method to engage more citizen-scientists. The elasticity of a cloud-based infrastructure also has potential to be applied to other natural hazards, such as droughts and landslides, at both global and regional scales

The limitations of the new reconstructed ET product are raised in the research. First, the reconstructed ET is a basin-mean product and correspondingly has variable spatial resolutions depending on the area of each individual sub-basin. Second, the ET reconstruction method does not account for the impacts of water transfer in or out of the sub-basins by human activities. However, the general similar spatial patterns between reconstructed ET and the other three ET products in these basins impacted by human activities suggest that most of these basins do not have substantial inter-basin transfer of water. Third, the ET estimates are calculated as the residual in water balance equation, which inherit the measurement and processing errors in all other water budget components. Finally, the availability of the ET reconstruction is limited by the availability of the observational measurements of the other water budget components. Further studies

can be conducted on the uncertainties of cross-basin water transfer and the error terms in water balance terms in ET calculation if related data are available. In addition, with different data source, it is possible to reconstruct actual ET estimates using the same methodology for basins lack of data in this study and perform further evaluation.

The web GIS-based hydrological modeling framework in this study is designed with capability to expand to other hydro-meteorological data with even higher spatial and temporal resolutions. For the model selection, only lumped hydrologic models are integrated in this framework. These are the tradeoffs among data availability, model complexity, and computation efficiency. The framework will eventually be applied at a global scale. Some regions in Africa do not have high-resolution data or lack of data for complicated models. It is wise to deploy the framework with relatively fine spatial resolution and simplified hydrologic models on our testbed with limited computation power. It is conceivable that the framework could be gradually migrated to high resolution with sophisticated hydrologic models in data rich areas, running on high performance clusters. Future research of this web GIS-based modeling framework will focus on three directions. First, the framework will be deployed on more powerful cluster to accommodate data with higher spatiotemporal resolutions in data rich areas, such as the CONUS. Evaluation will be performed on framework efficiency. Second, snow module for both lumped CREST and HyMOD models will be added and enabled, respectively. As a result, temperature data will be imported to the framework for snow module. Third, different types of models will be engaged and examined with this framework, including geotechnical model SLIDE. Distributed hydrologic models will also be evaluated with this web framework.

References

Adhikari, Pradeep, Yang Hong, Kimberly Douglas, DaliaBach Kirschbaum, Jonathan Gourley, Robert Adler, and Robert Brakenridge. 2010. "A digitized global flood inventory (1998–2008): compilation and preliminary results." *Natural hazards* 55 (2):405-422.

Adida, Ben. 1997. "It all starts at the server." *Internet Computing, IEEE* 1 (1):75-77.

Allen, Richard G, Luis S Pereira, Dirk Raes, and Martin Smith. 1998. "Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56." *FAO, Rome* 300 (9):D05109.

Anderson, Martha C, John M Norman, John R Mecikalski, Jason A Otkin, and William P Kustas. 2007. "A climatological study of evapotranspiration and moisture stress across the continental United States based on thermal remote sensing: 1. Model formulation." *Journal of Geophysical Research: Atmospheres (1984–2012)* 112 (D10).

Bastiaanssen, WGM, H Pelgrum, J Wang, Y Ma, JF Moreno, GJ Roerink, and T Van der Wal. 1998. "A remote sensing surface energy balance algorithm for land (SEBAL).: Part 2: Validation." *Journal of hydrology* 212:213-229.

Behzad, Babak, Anand Padmanabhan, Yong Liu, Yan Liu, and Shaowen Wang. 2011. "Integrating CyberGIS gateway with Windows Azure: a case study on MODFLOW groundwater simulation." Proceedings of the ACM SIGSPATIAL Second International Workshop on High Performance and Distributed Geographic Information Systems.

Bhatt, Gopal, Mukesh Kumar, and Christopher J. Duffy. 2014. "A tightly coupled GIS and distributed hydrologic modeling framework." *Environmental Modelling & Software* 62:70-84. doi: 10.1016/j.envsoft.2014.08.003.

Bhosale, Harshawardhan S, and Devendra P Gadekar. 2014. "A Review Paper on Big Data and Hadoop." 4 (10).

Borthakur, Dhruba, Jonathan Gray, Joydeep Sen Sarma, Kannan Muthukkaruppan, Nicolas Spiegelberg, Hairong Kuang, Karthik Ranganathan, Dmytro Molkov, Aravind Menon, and Samuel Rash. 2011. "Apache Hadoop goes realtime at

Facebook." Proceedings of the 2011 ACM SIGMOD International Conference on Management of data.

Cai, Ximing, Yi-Chen E Yang, Claudia Ringler, Jianshi Zhao, and Liangzhi You. 2011. "Agricultural water productivity assessment for the Yellow River Basin." *Agricultural water management* 98 (8):1297-1306.

Carpenter, Theresa M., and Konstantine P. Georgakakos. 2006. "Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales." *Journal of Hydrology* 329 (1-2):174-185. doi: 10.1016/j.jhydrol.2006.02.013.

Chambers, Don P. 2006. "Evaluation of new GRACE time‐variable gravity data over the ocean." *Geophysical Research Letters* 33 (17).

Chauhan, Seema, and RK Shrivastava. 2009. "Performance evaluation of reference evapotranspiration estimation using climate based methods and artificial neural networks." *Water resources management* 23 (5):825-837.

Chen, Fei, Kenneth Mitchell, John Schaake, Yongkang Xue, Hua-Lu Pan, Victor Koren, Qing Yun Duan, Michael Ek, and Alan Betts. 1996. "Modeling of land surface evaporation by four schemes and comparison with FIFE observations." *Journal of Geophysical Research. D. Atmospheres* 101:7251-7268.

Cleugh, Helen A, Ray Leuning, Qiaozhen Mu, and Steven W Running. 2007. "Regional evaporation estimates from flux tower and MODIS satellite data." *Remote Sensing of Environment* 106 (3):285-304.

Comair, Georges F., Daene C. McKinney, David R. Maidment, Gonzalo Espinoza, Harish Sangiredy, Abbas Fayad, and Fernando R. Salas. 2014. "Hydrology of the Jordan River Basin: A GIS-Based System to Better Guide Water Resources Management and Decision Making." *Water Resources Management* 28 (4):933-946. doi: 10.1007/s11269-014-0525-2.

Daly, Christopher, Michael Halbleib, Joseph I Smith, Wayne P Gibson, Matthew K Doggett, George H Taylor, Jan Curtis, and Phillip P Pasteris. 2008. "Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States." *International journal of climatology* 28 (15):2031.

DeVantier, Bruce A, and Arlen D Feldman. 1993. "Review of GIS applications in hydrologic modeling." *Journal of Water Resources Planning and Management* 119 (2):246-261.

Earth. 2015. "Earth Visualization." Accessed November 11. http://earth.nullschool.net/about.html.

Ek, MB, KE Mitchell, Y Lin, E Rogers, P Grunmann, V Koren, G Gayno, and JD Tarpley. 2003. "Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model." *Journal of Geophysical Research: Atmospheres (1984–2012)* 108 (D22).

ESRI. 2015. "ESRI ArcScene." Accessed November 11. http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/The_ArcScene_user_interface/00q8000000ws000000/.

Estellés-Arolas, Enrique, and Fernando González-Ladrón-de-Guevara. 2012. "Towards an integrated crowdsourcing definition." *Journal of Information science* 38 (2):189-200.

Fernández, Alberto, Sara del Río, Victoria López, Abdullah Bawakid, María J del Jesus, José M Benítez, and Francisco Herrera. 2014. "Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4 (5):380-409. doi: 10.1002/widm.1134.

Fisher, Joshua B, Kevin P Tu, and Dennis D Baldocchi. 2008. "Global estimates of the land–atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites." *Remote Sensing of Environment* 112 (3):901-919.

Fluet-Chouinard, Etienne, Bernhard Lehner, Lisa-Maria Rebelo, Fabrice Papa, and Stephen K Hamilton. 2015. "Development of a global inundation map at high spatial resolution from topographic downscaling of coarse-scale remote sensing data." *Remote Sensing of Environment* 158:348-361.

Freshwater Society. 2013. Minnesota's groundwater: Is our use sustainable? Excelsior, Minnesota: Freshwater Society.

Galletta, Dennis F, Raymond Henry, Scott McCoy, and Peter Polak. 2004. "Web site delays: How tolerant are users?" *Journal of the Association for Information Systems* 5 (1):1.

Gillies, RR, WP Kustas, and KS Humes. 1997. "A verification of the'triangle'method for obtaining surface soil water content and energy fluxes from remote measurements of the Normalized Difference Vegetation Index (NDVI) and surface e." *International journal of remote sensing* 18 (15):3145-3166.

Gong, Jianya, Peng Yue, and Hongxiu Zhou. 2010. "Geoprocessing in the Microsoft Cloud Computing Platform-Azure." Proceedings the Joint Symposium of ISPRS Technical Commission IV & AutoCarto.

Gonzalez, Hector, Alon Halevy, Christian S Jensen, Anno Langen, Jayant Madhavan, Rebecca Shapley, and Warren Shen. 2010. "Google fusion tables: data management, integration and collaboration in the cloud." Proceedings of the 1st ACM symposium on Cloud computing.

Goodchild, Michael F, and J Alan Glennon. 2010. "Crowdsourcing geographic information for disaster response: a research frontier." *International Journal of Digital Earth* 3 (3):231-241. doi: Doi 10.1080/17538941003759255.

Goodchild, Michael F, May Yuan, and Thomas J Cova. 2007. "Towards a general theory of geographic representation in GIS." *International journal of geographical information science* 21 (3):239-260.

Google. 2015a. "Google Earth." Accessed November 11. https://www.google.com/earth/.

Google. 2015b. "Google Fusion Table API." Accessed November 11. https://developers.google.com/fusiontables/.

Google. 2015c. "Google Maps JavaScript API." Accessed November 11. https://developers.google.com/maps/documentation/javascript/tutorial.

Gore, Al. 1998. "The digital earth: understanding our planet in the 21st century." *Australian surveyor* 43 (2):89-91.

Gourley, Jonathan J, Yang Hong, Zachary L Flamig, Ami Arthur, Robert Clark, Martin Calianno, Isabelle Ruin, Terry Ortel, Michael E Wieczorek, and Pierre-Emmanuel Kirstetter. 2013. "A unified flash flood database across the United States." *Bulletin of the American Meteorological Society* 94 (6):799-805.

Grayson, Rodger B., Ian D. Moore, and Thomas A. McMahon. 1992. "Physically based hydrologic modeling: 1. A terrain-based model for investigative purposes." *Water Resources Research* 28 (10):2639-2658. doi: 10.1029/92WR01258.

Herman, J. D., P. M. Reed, and T. Wagener. 2013. "Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior." *Water Resources Research* 49 (3):1400-1414. doi: 10.1002/wrcr.20124.

Herzog, Rolf. 1998. "PostgreSQL--The Linux of Databases." *Linux J.* 1998 (46es):1.

Hong, Yang, Xiaogang He, Amy Cerato, Ke Zhang, Zhen Hong, and Zonghu Liao. 2015. "Predictability of a Physically Based Model for Rainfall-induced Shallow Landslides: Model Development and Case Studies." In *Modern Technologies for Landslide Monitoring and Prediction*, edited by Marco Scaioni, 165-178. Springer Berlin Heidelberg.

Hu, Yao, Ximing Cai, and Benjamin DuPont. 2015. "Design of a web-based application of the coupled multi-agent system model and environmental model for watershed management analysis using Hadoop." *Environmental Modelling & Software* 70:149-162. doi: 10.1016/j.envsoft.2015.04.011.

Huai, Yin, Ashutosh Chauhan, Alan Gates, Gunther Hagleitner, Eric N. Hanson, Owen O'Malley, Jitendra Pandey, Yuan Yuan, Rubao Lee, and Xiaodong Zhang. 2014. "Major technical advancements in apache hive." Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, Snowbird, Utah, USA.

Huang, Bo. 2003. "Web-based dynamic and interactive environmental visualization." *Computers, Environment and Urban Systems* 27 (6):623-636. doi: 10.1016/s0198-9715(02)00063-7.

Huang, Qunying, Chaowei Yang, Karl Benedict, Songqing Chen, Abdelmounaam Rezgui, and Jibo Xie. 2013. "Utilize cloud computing to support dust storm forecasting." *International Journal of Digital Earth* 6 (4):338-355. doi: 10.1080/17538947.2012.749949.

Hudson-Smith, Andrew, Michael Batty, Andrew Crooks, and Richard Milton. 2009. "Mapping for the masses accessing Web 2.0 through crowdsourcing." *Social Science Computer Review* 27 (4):524-538. doi: 10.1177/0894439309332299.

Jung, Martin, Markus Reichstein, Philippe Ciais, Sonia I Seneviratne, Justin Sheffield, Michael L Goulden, Gordon Bonan, Alessandro Cescatti, Jiquan Chen, and Richard De Jeu. 2010. "Recent decline in the global land evapotranspiration trend due to limited moisture supply." *Nature* 467 (7318):951-954.

Khakbaz, Behnaz, Bisher Imam, Kuolin Hsu, and Soroosh Sorooshian. 2012. "From lumped to distributed via semi-distributed: Calibration strategies for semi-distributed hydrologic models." *Journal of Hydrology* 418-419:61-77. doi: 10.1016/j.jhydrol.2009.02.021.

Khan, Sadiq Ibrahim, Yang Hong, Baxter Vieux, and Wenjuan Liu. 2010. "Developmentevaluation of an actual evapotranspiration estimation algorithm using satellite remote sensingmeteorological observational network in Oklahoma." *International Journal of Remote Sensing* 31 (14):3799-3819.

Kitchin, Rob. 2013. "Big data and human geography Opportunities, challenges and risks." *Dialogues in Human Geography* 3 (3):262-267.

Koren, V, J Schaake, K Mitchell, Q‐Y Duan, F Chen, and JM Baker. 1999. "A parameterization of snowpack and frozen ground intended for NCEP weather and climate models." *Journal of Geophysical Research: Atmospheres (1984–2012)* 104 (D16):19569-19585.

Koren, Victor, Michael Smith, Zhengtao Cui, and Brian Cosgrove. 2007. NOAA Technical Report NWS 52: Physically-based modifications to the Sacramento Soil Moisture Account Model: Modeling the effects of frozen ground on the rainfall-runoff process. edited by NOAA National Weather Service Office of Hydrologic Development. Silver Spring, Maryland.

Koster, Randal D, Max J Suarez, and Mark Heiser. 2000. "Variance and predictability of precipitation at seasonal-to-interannual timescales." *Journal of hydrometeorology* 1 (1):26-46.

Kussul, Nataliia, Dan Mandl, Karen Moe, J-P Mund, Joachim Post, Andrii Shelestov, Sergii Skakun, Joerg Szarzynski, Guido Van Langenhove, and Matthew Handy. 2012. "Interoperable infrastructure for flood monitoring: Sensorweb, Grid and

cloud." *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of* 5 (6):1740 - 1745. doi: 10.1109/JSTARS.2012.2192417.

Landerer, FW, and SC Swenson. 2012. "Accuracy of scaled GRACE terrestrial water storage estimates." *Water Resources Research* 48 (4).

Leaflet. 2015. "Leaflet open-source JavaScript library." Accessed November 11. http://leafletjs.com/.

Leavesley, George H. 1994. "Modeling the effects of climate change on water resources-a review." *Climatic Change* 28 (1-2):159-177.

Li, Weiyue, Chun Liu, Zhanming Wan, Yang Hong, Weiwei Sun, Zhiwei Jian, and Sheng Chen. 2013. "A Cloud-based China's Landslide Disaster Database (CCLDD)."

Li, Zhenlong, Chaowei Yang, Min Sun, Jing Li, Chen Xu, Qunying Huang, and Kai Liu. 2013. "A high performance web-based system for analyzing and visualizing spatiotemporal data for climate studies." In *Web and Wireless Geographical Information Systems*, 190-198. Springer.

Liang, Xetal, Dennis P Lettenmaier, Eric F Wood, and Stephen J Burges. 1994. "A simple hydrologically based model of land surface water and energy fluxes for general circulation models." *JOURNAL OF GEOPHYSICAL RESEARCH-ALL SERIES*- 99:14,415-14,415.

Liang, Xu, Eric F Wood, and Dennis P Lettenmaier. 1996. "Surface soil moisture parameterization of the VIC-2L model: Evaluation and modification." *Global and Planetary Change* 13 (1):195-206.

Lim, Kyoung Jae, Bernard A Engel, Zhenxu Tang, Joongdae Choi, Ki‐Sung Kim, Suresh Muthukrishnan, and Dibyajyoti Tripathy. 2005. "Automated web gis based hydrograph analysis tool, WHAT." *Journal of the American Water Resources Association* 41 (6):1407–1416.

Liu, Wenjuan, Yang Hong, Sadiq Khan, Mingbin Huang, Trevor Grout, and Pradeep Adhikari. 2011. "Evaluation of Global Daily Reference ET Using Oklahoma's Environmental Monitoring Network—MESONET." *Water resources management* 25 (6):1601-1613.

Liu, Yuqiong, and Hoshin V. Gupta. 2007. "Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework." *Water Resources Research* 43 (7):n/a-n/a. doi: 10.1029/2006wr005756.

Long, Di, Laurent Longuevergne, and Bridget R Scanlon. 2014. "Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites." *Water Resources Research* 50 (2):1131-1151.

Long, Di, Bridget R Scanlon, Laurent Longuevergne, Alexander Y Sun, D Nelun Fernando, and Himanshu Save. 2013. "GRACE satellite monitoring of large depletion in water storage in response to the 2011 drought in Texas." *Geophysical Research Letters* 40 (13):3395-3401.

Longuevergne, Laurent, Bridget R Scanlon, and Clark R Wilson. 2010. "GRACE Hydrological estimates for small basins: Evaluating processing approaches on the High Plains Aquifer, USA." *Water Resources Research* 46 (11).

Mallick, Kaniska, Andrew J Jarvis, Eva Boegh, Joshua B Fisher, Darren T Drewry, Kevin P Tu, Simon J Hook, Glynn Hulley, Jonas Ardö, and Jason Beringer. 2014. "A Surface Temperature Initiated Closure (STIC) for surface energy balance fluxes." *Remote Sensing of Environment* 141:243-261.

Mantas, V. M., Z. Liu, and A. J. S. C. Pereira. 2015. "A web service and android application for the distribution of rainfall estimates and Earth observation data." *Computers & Geosciences* 77:66-76. doi: 10.1016/j.cageo.2015.01.011.

Martínez-López, Javier, María F Carreño, José A Palazón-Ferrando, Julia Martínez-Fernández, and Miguel A Esteve. 2014. "Free advanced modeling and remote-sensing techniques for wetland watershed delineation and monitoring." *International Journal of Geographical Information Science* 28 (8):1610-1625.

Mattikalli, NM. 1995. "Integration of remotely-sensed raster data with a vector-based geographical information system for land-use change detection." *Remote Sensing* 16 (15):2813-2828.

McCuen, Richard H. 1973. "The role of sensitivity analysis in hydrologic modeling." *Journal of Hydrology* 18 (1):37-53.

Mell, Peter, and Timothy Grance. 2011. "The NIST definition of cloud computing (draft)." *NIST special publication* 800:145.

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. 1953. "Equation of state calculations by fast computing machines." *The journal of chemical physics* 21 (6):1087-1092.

Mitchell, Kenneth E, Dag Lohmann, Paul R Houser, Eric F Wood, John C Schaake, Alan Robock, Brian A Cosgrove, Justin Sheffield, Qingyun Duan, and Lifeng Luo. 2004. "The multi‐institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system." *Journal of Geophysical Research: Atmospheres (1984–2012)* 109 (D7).

Monteith, J. L. 1965. "Evaporation and environment." In *The state and movement of water in living organisms, Symposium of the society of experimental biology*, 205-234. Cambridge: Cambridge University Press.

Moore, RJ. 2007. "The PDM rainfall-runoff model." *Hydrology and Earth System Sciences Discussions* 11 (1):483-499.

Mu, Qiaozhen, Faith Ann Heinsch, Maosheng Zhao, and Steven W Running. 2007. "Development of a global evapotranspiration algorithm based on MODIS and global meteorology data." *Remote Sensing of Environment* 111 (4):519-536.

Mu, Qiaozhen, Maosheng Zhao, and Steven W Running. 2011. "Improvements to a MODIS global terrestrial evapotranspiration algorithm." *Remote Sensing of Environment* 115 (8):1781-1800.

Nishida, Kenlo, Ramakrishna R Nemani, Joseph M Glassy, and Steven W Running. 2003. "Development of an evapotranspiration index from Aqua/MODIS for monitoring surface moisture status." *Geoscience and Remote Sensing, IEEE Transactions on* 41 (2):493-501.

ODK. 2015. "Open Data Kit." Accessed November 11. https://opendatakit.org/.

Olivera, Francisco, Milver Valenzuela, R Srinivasan, Janghwoan Choi, Hiudae Cho, Srikanth Koka, and Ashish Agrawal. 2006. "ArcGIS-SWAT: A geodata model and GIS interface for SWAT." *Journal of the American Water Resources Association* 42 (2):295-309.

Polato, Ivanilton, Reginaldo Ré, Alfredo Goldman, and Fabio Kon. 2014. "A comprehensive view of Hadoop research—A systematic literature review."

Journal of Network and Computer Applications 46:1-25. doi: 10.1016/j.jnca.2014.07.022.

Quan, Z., J. Teng, W. Sun, T. Cheng, and J. Zhang. 2015. "Evaluation of the HyMOD model for rainfall–runoff simulation using the GLUE method." *Proceedings of the International Association of Hydrological Sciences* 368:180-185. doi: 10.5194/piahs-368-180-2015.

Ramillien, Guillaume, Frédéric Frappart, Andreas Güntner, Thanh Ngo‐Duc, Anny Cazenave, and Katia Laval. 2006. "Time variations of the regional evapotranspiration rate from Gravity Recovery and Climate Experiment (GRACE) satellite gravimetry." *Water Resources Research* 42 (10).

Reed, Seann, Victor Koren, Michael Smith, Ziya Zhang, Fekadu Moreda, Dong-Jun Seo, and and Dmip Participants. 2004. "Overall distributed model intercomparison project results." *Journal of Hydrology* 298 (1-4):27-60. doi: 10.1016/j.jhydrol.2004.03.031.

Rossum, Guido van. 1997. "Scripting the web with python." *World Wide Web Journal* 2 (2):97-120.

Sanner, Michel F. 1999. "Python: a programming language for software integration and development." *J Mol Graph Model* 17 (1):57-61.

Schwalm, Christopher R, Deborah N Huntinzger, Anna M Michalak, Joshua B Fisher, John S Kimball, Brigitte Mueller, Ke Zhang, and Yongqiang Zhang. 2013. "Sensitivity of inferred climate model skill to evaluation decisions: a case study using CMIP5 evapotranspiration." *Environmental Research Letters* 8 (2):024028.

Shvachko, Konstantin, Hairong Kuang, Sanjay Radia, and Robert Chansler. 2010. "The hadoop distributed file system." Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on.

Steiniger, Stefan, and Andrew JS Hunter. 2013. "The 2012 free and open source GIS software map–A guide to facilitate research, development, and adoption." *Computers, Environment and Urban Systems* 39:136-150.

Su, Z. 2002. "The Surface Energy Balance System (SEBS) for estimation of turbulent heat fluxes." *Hydrology and Earth System Sciences Discussions* 6 (1):85-100.

Sun, Alexander. 2013. "Enabling collaborative decision-making in watershed management using cloud-computing services." *Environmental Modelling & Software* 41:93-97. doi: 10.1016/j.envsoft.2012.11.008.

Swain, Nathan R, Kilisimasi Latu, Scott D Christensen, Norman L Jones, E James Nelson, Daniel P Ames, and Gustavious P Williams. 2015. "A review of open source software solutions for developing water resources web applications." *Environmental Modelling & Software* 67:108-117.

Swenson, Sean, and John Wahr. 2006. "Post-processing removal of correlated errors in GRACE data." *Geophysical Research Letters* 33 (8):L08402. doi: 10.1029/2005GL025285.

Tang, Qiuhong, Shannon Peterson, Richard H Cuenca, Yutaka Hagimoto, and Dennis P Lettenmaier. 2009. "Satellite ‐ based near ‐ real ‐ time estimation of irrigated crop water consumption." *Journal of Geophysical Research: Atmospheres (1984 ‐2012)* 114 (D5).

Tapley, Byron D, S Bettadpur, M Watkins, and Ch Reigber. 2004. "The gravity recovery and climate experiment: Mission overview and early results." *Geophysical Research Letters* 31 (9).

Tapley, Byron D, Srinivas Bettadpur, John C Ries, Paul F Thompson, and Michael M Watkins. 2004. "GRACE measurements of mass variability in the Earth system." *Science* 305 (5683):503-505.

Thusoo, Ashish, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu, and Raghotham Murthy. 2010. "Hive-a petabyte scale data warehouse using hadoop." Data Engineering (ICDE), 2010 IEEE 26th International Conference on.

V. Boyina, Ramana Prasad, Glenn P. Catts, Charlynne T. Smith, and Hugh A. Devine. 2015. "Hydrologic Web-Mapping Application of Hofmann Forest with GIS Approach: Case Study." *Journal of Hydrologic Engineering*:E5015006. doi: 10.1061/(asce)he.1943-5584.0001285.

Van Der Knijff, J. M., J. Younis, and A. P. J. De Roo. 2010. "LISFLOOD: a GIS ‐ based distributed model for river basin scale water balance and flood simulation." *International Journal of Geographical Information Science* 24 (2):189-212. doi: 10.1080/13658810802549154.

Velpuri, Naga M, Gabriel B Senay, Ramesh K Singh, Stefanie Bohms, and James P Verdin. 2013. "A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET." *Remote Sensing of Environment* 139:35-49.

Viessman, W., and G.L. Lewis. 2003. *Introduction to Hydrology*: Prentice Hall.

Vieux, Baxter. 2001. "Distributed Hydrologic Modeling Using GIS." In *Distributed Hydrologic Modeling Using GIS*, 1-17. Springer Netherlands.

Vitolo, Claudia, Yehia Elkhatib, Dominik Reusser, Christopher J. A. Macleod, and Wouter Buytaert. 2015. "Web technologies for environmental Big Data." *Environmental Modelling & Software* 63:185-198. doi: 10.1016/j.envsoft.2014.10.007.

Wagener, Thorsten, Douglas P Boyle, Matthew J Lees, Howard S Wheater, Hoshin V Gupta, and Soroosh Sorooshian. 2001. "A framework for development and application of hydrological models." *Hydrology and Earth System Sciences Discussions* 5 (1):13-26.

Wahr, John, Mery Molenaar, and Frank Bryan. 1998. "Time variability of the Earth's gravity field: Hydrological and oceanic effects and their possible detection using GRACE." *Journal of Geophysical Research: Solid Earth (1978–2012)* 103 (B12):30205-30229.

Wahr, John, Sean Swenson, and Isabella Velicogna. 2006. "Accuracy of GRACE mass estimates." *Geophysical Research Letters* 33 (6).

Wan, Zhanming, Yang Hong, Sadiq Khan, Jonathan Gourley, Zachary Flamig, Dalia Kirschbaum, and Guoqiang Tang. 2014. "A cloud-based global flood disaster community cyber-infrastructure: Development and demonstration." *Environmental Modelling & Software* 58:86-94. doi: DOI 10.1016/j.envsoft.2014.04.007.

Wan, Zhanming, Ke Zhang, Xianwu Xue, Zhen Hong, Yang Hong, and Jonathan J Gourley. 2015. "Water balance‐based actual evapotranspiration reconstruction from ground and satellite observations over the conterminous United States." *Water Resources Research* 51 (8):6485-6499.

Wang, Dingbao, and Negin Alimohammadi. 2012. "Responses of annual runoff, evaporation, and storage change to climate variability at the watershed scale." *Water Resources Research* 48 (5).

Wang, J, and Rafael L Bras. 2009. "A model of surface heat fluxes based on the theory of maximum entropy production." *Water resources research* 45 (11).

Wang, Jiahu, Yang Hong, Li Li, Jonathan J. Gourley, Sadiq I. Khan, Koray K. Yilmaz, Robert F. Adler, Frederick S. Policelli, Shahid Habib, Daniel Irwn, Ashutosh S. Limaye, Tesfaye Korme, and Lawrence Okello. 2011. "The coupled routing and excess storage (CREST) distributed hydrological model." *Hydrological Sciences Journal* 56 (1):84-98. doi: 10.1080/02626667.2010.543087.

Wang, Jingfeng, and RL Bras. 2011. "A model of evapotranspiration based on the theory of maximum entropy production." *Water Resources Research* 47 (3).

Wilson, Tim. 2008. *OGC KML, Version 2.2. 0*: Open Geospatial Consortium.

Wood, Eric F, Dennis Lettenmaier, Xu Liang, Bart Nijssen, and Suzanne W Wetzel. 1997. "Hydrological modeling of continental-scale basins." *Annual Review of Earth and Planetary Sciences* 25 (1):279-300.

Wu, Huan, John S Kimball, Hongyi Li, Maoyi Huang, L Ruby Leung, and Robert F Adler. 2012. "A new global river network database for macroscale hydrologic modeling." *Water resources research* 48 (9).

Xia, Youlong, Michael T Hobbins, Qiaozhen Mu, and Michael B Ek. 2015. "Evaluation of NLDAS‐2 evapotranspiration against tower flux site observations." *Hydrological Processes* 29 (7):1757-1771.

Xia, Youlong, Kenneth Mitchell, Michael Ek, Brian Cosgrove, Justin Sheffield, Lifeng Luo, Charles Alonge, Helin Wei, Jesse Meng, and Ben Livneh. 2012. "Continental‐scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS‐2): 2. Validation of model‐simulated streamflow." *Journal of Geophysical Research: Atmospheres (1984–2012)* 117 (D3).

Xia, Youlong, Kenneth Mitchell, Michael Ek, Justin Sheffield, Brian Cosgrove, Eric Wood, Lifeng Luo, Charles Alonge, Helin Wei, and Jesse Meng. 2012.

"Continental‐scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS‐2): 1. Intercomparison and application of model products." *Journal of Geophysical Research: Atmospheres (1984–2012)* 117 (D3).

Xue, Xianwu, Yang Hong, Ashutosh S. Limaye, Jonathan J. Gourley, George J. Huffman, Sadiq Ibrahim Khan, Chhimi Dorji, and Sheng Chen. 2013. "Statistical and hydrological evaluation of TRMM-based Multi-satellite Precipitation Analysis over the Wangchu Basin of Bhutan: Are the latest satellite precipitation products 3B42V7 ready for use in ungauged basins?" *Journal of Hydrology* 499:91-99. doi: 10.1016/j.jhydrol.2013.06.042.

Xue, Xianwu, Ke Zhang, Yang Hong, Jonathan J. Gourley, Wayne Kellogg, Renee A. McPherson, Zhanming Wan, and Barney N. Austin. 2015. "New Multisite Cascading Calibration Approach for Hydrological Models: Case Study in the Red River Basin Using the VIC Model." *Journal of Hydrologic Engineering*:05015019. doi: 10.1061/(asce)he.1943-5584.0001282.

Yang, Chaowei, Michael Goodchild, Qunying Huang, Doug Nebert, Robert Raskin, Yan Xu, Myra Bambacus, and Daniel Fay. 2011. "Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing?" *International Journal of Digital Earth* 4 (4):305-329.

Zavala-Romero, Olmo, Arsalan Ahmed, Eric P Chassignet, Jorge Zavala-Hidalgo, Agustin Fernández Eguiarte, and Anke Meyer-Baese. 2014. "An open source Java web application to build self-contained web GIS sites." *Environmental Modelling & Software* 62:210-220.

Zeng, Zhenzhong, Shilong Piao, Xin Lin, Guodong Yin, Shushi Peng, Philippe Ciais, and Ranga B Myneni. 2012. "Global evapotranspiration over the past three decades: estimation based on the water balance equation combined with empirical models." *Environmental Research Letters* 7 (1):014026.

Zetter, Roger. 2012. *World disaster report: focus on forced migration and displacement*: International Federation of Red Cross and Red Crescent Societies.

Zhan, Xiaoyong, and Min-Lang Huang. 2004. "ArcCN-Runoff: an ArcGIS tool for generating curve number and runoff maps." *Environmental Modelling & Software* 19 (10):875-879.

Zhang, Ke, John S Kimball, Ramakrishna R Nemani, and Steven W Running. 2010. "A continuous satellite‐derived global record of land surface evapotranspiration from 1983 to 2006." *Water Resources Research* 46 (9).

Zhang, Ke, John S Kimball, Ramakrishna R Nemani, Steven W Running, Yang Hong, Jonathan J Gourley, and Zhongbo Yu. 2015. "Vegetation Greening and Climate Change Promote Multidecadal Rises of Global Land Evapotranspiration." *Scientific reports* 5.

Zhang, Ke, John S. Kimball, Qiaozhen Mu, Lucas A. Jones, Scott J. Goetz, and Steven W. Running. 2009. "Satellite based analysis of northern ET trends and associated changes in the regional water balance from 1983 to 2005." *Journal of Hydrology* 379 (1):92-110. doi: 10.1016/j.jhydrol.2009.09.047.

Zhang, YQ, FHS Chiew, L Zhang, R Leuning, and HA Cleugh. 2008. "Estimating catchment evaporation and runoff using MODIS leaf area index and the Penman‐Monteith equation." *Water Resources Research* 44 (10).

Zhang, Yu, Yang Hong, Jonathan J Gourley, Xuguang Wang, G Robert Brakenridge, Tom De Groeve, and Humberto Vergara. 2014. "Impact of Assimilating Spaceborne Microwave Signals for Improving Hydrological Prediction in Ungauged Basins." *Remote Sensing of the Terrestrial Water Cycle* 206:439.

Zhang, Yu, Yang Hong, XuGuang Wang, Jonathan J Gourley, JiDong Gao, Humberto J Vergara, and Bin Yong. 2013. "Assimilation of passive microwave streamflow signals for improving flood forecasting: A first study in Cubango river basin, Africa." *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of* 6 (6):2375-2390.

Zhao, R.-J., Y.-L. Zhang, L.-R. Fang, X.-R. Liu, and Q.-S. Zhang. 1980. "The Xinanjiang Model." In *Hydrological Forecasting Proceedings Oxford Symposium*, 351-356. IAHS Press.

Zhao, Ren-Jun. 1992. "The Xinanjiang model applied in China." *Journal of Hydrology* 135 (1):371-381.